Journal of Machine Learning Research 1 (2016) 1-9999 Submitted 4/00; Published 10/00

Bipartite ranking: a risk-theoretic perspective

Aditya Krishna Menon ADITYA.MENON@DATA61.CSIRO.AU
Robert C. Williamson BOB.WILLIAMSON@ANU.EDU.AU
Data61 and the Australian National University
Canberra, ACT, Australia

Editor: Nicolas Vayatis

Abstract

We present a systematic study of the bipartite ranking problem, with the aim of explicating its connections to the class-probability estimation problem. Our study focuses on the properties of the statistical risk for bipartite ranking with general losses, which is closely related to a generalised notion of the area under the ROC curve: we establish alternate representations of this risk, relate the Bayes-optimal risk to a class of probability divergences, and characterise the set of Bayes-optimal scorers for the risk. We further study properties of a generalised class of bipartite risks, based on the p-norm push of Rudin (2009). Our analysis is based on the rich framework of proper losses, which are the central tool in the study of class-probability estimation. We show how this analytic tool makes transparent the generalisations of several existing results, such as the equivalence of the minimisers for four seemingly disparate risks from bipartite ranking and class-probability estimation. A novel practical implication of our analysis is the design of new families of losses for scenarios where accuracy at the head of ranked list is paramount, with comparable empirical performance to the p-norm push.

Keywords: Bipartite ranking, class-probability estimation, proper losses, Bayes-optimality, ranking the best

1. The bipartite ranking problem

Bipartite ranking problems (Freund et al., 2003; Agarwal et al., 2005; Clémençon et al., 2008; Kotlowski et al., 2011) have received considerable attention from the machine learning community. In such problems, we have as input a training set of examples, each of which comprises an *instance* (typically a vector of features describing some entity) with an associated *binary label* (typically denoted "positive" or "negative", describing whether the instance possesses some attribute). The goal is to learn a *scorer*, which assigns to each instance a real number, such that positive instances have a higher score than negative instances. Violations of this condition are penalised according to some loss ℓ, and the *bipartite ranking risk* of a scorer is its expected penalty according to ℓ. When ℓ corresponds to the 0-1 loss, the bipartite ranking risk is one minus the *area under the ROC curve* (AUC) of the scorer (Agarwal and Niyogi, 2005; Clémençon et al., 2008; Krzanowski and Hand, 2009). Applications of bipartite ranking range from content recommendation, where the goal is to rank a set of items based on an individual's preference for them, to epidemiological studies, where the goal is to a rank a set of individuals based on how likely they are to have a particular disease.

While bipartite ranking has received considerable study, the focus has primarily been on *algorithm design*. There has been relatively little theoretical study of issues such as the properties of its statistical risk, and it is only recently that its relationship to extant supervised learning problems has been formally established (Narasimhan and Agarwal, 2013b). While the design of computationally and statistically efficient methods for bipartite ranking is important, we believe there is value in explicating the statistical risk assumed by the problem, the optimal solutions that result from it, and the implied relationships to other learning problems.

To this end, in this paper[1], we systematically study bipartite ranking through its statistical risk. In brief, we study the properties of the bipartite risk (and hence the AUC) for an arbitrary scorer, the properties of the Bayes-optimal bipartite risk and the bipartite regret for an arbitrary scorer, and characterise the set of

1. A preliminary version of this work appeared in (Menon and Williamson, 2014).

the Bayes-optimal scorers. While some of these topics have been touched upon in prior studies, we aim to provide a comprehensive, unified treatment of the material. Our analysis rests heavily upon the framework of proper losses (Buja et al., 2005; Reid and Williamson, 2010) – the machinery underlying the analysis of the class-probability estimation problem – which we hope to demonstrate to be the natural lens with which to study bipartite ranking problems. The proper loss framework has previously been employed in the analysis of a reduction of bipartite ranking to class-probability estimation (Agarwal, 2014). In this paper, we show how this framework additionally provides a clean way of generalising existing results on the Bayes-optimal scorers (§7.3, §9.5), makes transparent the connections between bipartite ranking and class-probability estimation (§10.2), and immediately establishes the equivalence of minimisers for seemingly disparate risks (§11). A novel practical implication is a means of designing losses suitable for the task of "ranking the best" (§9.6), which we show to perform favourably compared with existing approaches (§9.7).

Table 1 provides an overview of the material covered in this paper. In more detail:

- We formally define the bipartite ranking problem for a general loss via its statistical risk (§3.3), and derive its equivalence to a classification problem over pairs (§4).
- We study the properties of the ROC curve, such as its connection to the calibration transform (§5.2.5), and its value in determining thresholds for cost-sensitive classification (§5.2.6). We derive a (to our knowledge novel) result (Proposition 13) on how dominance of one calibrated scorer over another in ROC space implies dominance with respect to *any* proper composite loss, which establishes the coherence of using the ROC curve to compare calibrated scorers.
- We discuss several interpretation of the AUC, including its relationship to the bipartite risk (§5.5) and a number of integral representations (§5.6). We show how one of these representations, due to Hand (2009), is related to the integral representation for proper losses, and discuss its implications for the coherence of the use of AUC to compare scorers (§5.6.2).
- We relate the Bayes-optimal bipartite risk to an f-divergence between product measures for the class-conditional distributions (§6.2), generalising a result for the case of 0-1 loss due to Torgersen (1991). We further relate the bipartite regret to a generative Bregman divergence (§6.3).
- We determine the set of Bayes-optimal scorers for surrogate bipartite ranking risks (§7.3, §7.4, §7.5), demonstrating how the proper loss framework helps generalise existing results on the topic. We use these results to derive surrogate regret bounds, and thus AUC-consistency, for algorithms that minimise a suitable surrogate loss over pairs (§8).
- We formalise the "ranking the best" extension to bipartite ranking (§9.1), and the study the Bayes-optimal scorers for the p-norm push risk (§9.5). We show how the risk can be related to a proper composite loss with asymmetric weight function over misclassification costs (§9.6.1).
- We show how the weight function view of a proper composite loss suggests a strategy for designing losses suitable for "ranking the best". We then describe several such new loss functions (§9.6.3). We evaluate these losses empirically on a number of real-world datasets (§9.7), and demonstrate their favourable empirical performance compared to the p-norm push risk.
- Based on the corresponding Bayes-optimal solutions, we relate bipartite ranking to the learning problems of pairwise ranking, class-probability estimation, and classification (§10.1, §10.2). This formally elucidates the relative "difficulty" of each of these problems.
- Based on the corresponding Bayes-optimal solutions, we establish the equivalence between the minimisers of seemingly disparate risks for four popular approaches to bipartite ranking (§11). This further illustrates the close links between bipartite ranking and class-probability estimation.
- We relate bipartite ranking to axiomatic characterisations of ranking relations, and in particular show how theorems characterising the existence of utility representations describe the class of ranking problems over pairs that it can model (§12.4).

Topic	Description	Reference
RISK	Bipartite ranking risk for general loss ℓ	§3.3
	Equivalence to pairwise ranking risk	§4
RELATION TO AUC	AUC and generalisation to a general loss ℓ	§5.3, §5.4
	Equivalence of bipartite ranking risk and ℓ-AUC	§5.5
	Integral representations of the AUC and ℓ-AUC	§5.6
	AUC and the Neyman-Pearson problem	§D
OPTIMAL RISK	Relationship between Bayes risk and f-divergences	§6.2
	Relationship between regret and generative Bregman divergences	§6.3
OPTIMAL SCORERS	Bayes-optimal pair-scorers	§7.3
	Bayes-optimal univariate scorers	§7.4, §7.5
SURROGATE REGRET	Surrogate regret bounds for pairwise minimisation	§8
GENERALISED RISK	Ranking the best formulation	§9.1
	Bayes-optimal scorers for p-norm push	§9.5
	Proper composite approach to ranking the best	§9.6
	Empirical comparison of algorithms for ranking the best	§9.7
EQUIVALENCES	Reduction to classification and class-probability estimation	§10.1, §10.2
	Equivalent risks for bipartite ranking	§11
AXIOMATIC CHARACTERISATION	Utility representation theorems	§12.4

Table 1: Summary of the results on bipartite ranking in this paper.

Before initiating our study with a description of bipartite risk, we fix notation and provide definitions of key quantities that will be used throughout the paper.

2. Preliminary definitions and notation

We define the relevant quantities used in the rest of the paper, and fix some notation. Table 2 provides a glossary of some frequently used symbols.

2.1 Notation

We use scripted calligraphic fonts e.g. $\mathcal{X}, \mathcal{Y}$ to denote sets. We use $\mathcal{X} \setminus \mathcal{Y}$ to denote set difference, and $\emptyset$ to denote the empty set.

We denote by $\mathbb{R}$ the set of real numbers, and $\mathbb{R}_+ = [0, \infty)$. For a positive integer n, we write $[n] = \{1, 2, \ldots, n\}$. For a function $f : \mathcal{X} \to \mathbb{R}$, we denote its image or range by $\mathrm{Im}(f)$. If f is differentiable, we denote its derivative by f'. For functions f, g, $f \circ g$ denotes functional composition, so that $(f \circ g)(x) = f(g(x))$. For a nonincreasing function $f : \mathbb{R} \to \mathbb{R}$, define the pseudo-inverse $f^{-1} : \mathbb{R} \to \mathbb{R}$ by

$$f^{-1}(y) \doteq \inf\{x \in \mathbb{R} : f(x) \leq y\}. \tag{1}$$

If f is strictly decreasing, this coincides with the standard inverse function. When f is nondecreasing, we replace the inf with a sup in Equation 1. For a constant $c \in \mathbb{R}$, we write $f \equiv c$ to mean that $f(x) = c$ for every $x \in \mathcal{X}$.

Symbol	Meaning	Definition
$[\![\cdot]\!]$	Indicator function	§2.1
$\wedge, \vee$	Minimum and maximum	§2.1
$\sigma(\cdot)$	Sigmoid function	§2.1
$\mathcal{X}$	Instance space, typically $\mathbb{R}^n$	§2.1
$\mathcal{Y}$	Label space, typically $\{\pm 1\}$	§2.1
$\Delta_{\mathcal{S}}$	Set of all distributions over a set $\mathcal{S}$	§2.2
X, Y	Random variables, typically samples from D	§2.2
D	Distribution over $\mathcal{X} \times \{\pm 1\}$	§3.1
R	Distribution over $\mathcal{X} \times \mathcal{X} \times \{\pm 1\}$	§3.4
D_{BR}	Pairwise ranking distribution derived from D	§3.3
P, Q, p, q	Class-conditional distributions and densities of D	§3.2
M, μ	Observation distribution and density of D	§3.2
η	Observation-conditional distribution of D	§3.2
η_{Pair}	Observation-conditional distribution of D_{BR}	§3.3
π	Positive class base rate of D	§3.2
$D = \langle P, Q, \pi \rangle = \langle M, \eta \rangle$	Constituent components of distribution D	§3.1
$s : \mathcal{X} \to \mathbb{R}$	Scorer	§2.4
$s_{\text{Pair}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$	Pair-scorer	§2.4
$\text{Diff}(s) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$	Difference pair-scorer	§2.4
$\mathcal{S}_{\text{Decomp}}$	Set of all decomposable pair-scorers	§2.4
$\text{Prb}(\cdot; D, s) : \mathbb{R} \to [0, 1]$	Score-to-probability transform	Equation 5
$\text{Cal}(\cdot; D, s) : \mathcal{X} \to [0, 1]$	Calibration transform	Equation 5
S, S	Random variable and distribution of scores	§2.4
$\ell(\cdot, \cdot)$	Loss function	§2.6
$\ell_{\text{symm}}(\cdot, \cdot)$	Symmetrised loss function	Equation 8
$\Psi(\cdot)$	Link function for a proper composite loss	§2.6
$w(\cdot)$	Weight function for a proper composite loss	§2.6
$\mathcal{L}_{\text{SPC}}$	Set of all strictly proper composite losses	§2.6
$\mathcal{L}_{\text{SPC}}(\Psi)$	Set of all strictly proper composite losses with link Ψ	§2.6
$\mathcal{L}_{\text{Decomp}}$	Set of all losses with decomposable Bayes-optimal pair-scorer	Equation 20
$L(\cdot, \cdot; \ell)$	Conditional risk for loss ℓ	§3.2
$L^*(\cdot; \ell)$	Bayes-optimal conditional risk for loss ℓ	§3.2
$\mathbb{L}(\cdot; D, \ell)$	Risk for loss ℓ	§3.2
$\mathbb{L}_{\text{BR}}(\cdot; D, \ell)$	Bipartite risk for loss ℓ	§3.3
$\mathbb{L}^*(D, \ell)$	Bayes-optimal (minimal) risk for loss ℓ	§3.2
$\mathbb{L}^*_{\text{BR}}(D, \ell)$	Bayes-optimal (minimal) bipartite risk over scorers for loss ℓ	§3.3
$\mathcal{S}^*(D, \ell)$	Set of Bayes-optimal scorers for classification for loss ℓ	§3.2
$\mathcal{S}^*_{\text{BR}}(D, \ell)$	Set of Bayes-optimal scorers for bipartite ranking for loss ℓ	§3.3
$\text{regret}(\cdot; D, \ell)$	Classification regret of scorer for loss ℓ	§3.2
$\text{regret}_{\text{BR}}(\cdot; D, \ell)$	Bipartite ranking regret of scorer for loss ℓ	§3.3
$\mathbb{I}_f(\cdot, \cdot)$	f-divergence between distributions	Equation 3
$\mathbb{B}_f(\cdot, \cdot)$	Generative Bregman divergence between distributions	Equation 4
TPR, TNR, FPR, FNR	True (false) positive (negative) rates	Definition 4
AUC, AUC_ℓ	Area under the ROC curve, 0-1 and ℓ loss	§5.3, §5.4

Table 2: Glossary of frequently used symbols used in this paper.

For any function $f : \mathcal{X} \to \mathbb{R}$, we denote by $\operatorname{Argmin}_{x\in\mathcal{X}} f(x)$ the set of all minimisers i.e. all $x \in \mathcal{X}$ such that $f(x) \leq f(x')$ for all $x' \in \mathcal{X}$. When the set is a singleton, so that f has a unique minimiser, we denote this minimiser by $\operatorname{argmin}_{x\in\mathcal{X}} f(x)$.

We denote by $\operatorname{Diff}(f) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ the function satisfying $(\operatorname{Diff}(f))(x, x') = f(x) - f(x')$ for every $x, x' \in \mathcal{X}$. For a set of functions $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}\}$, we define $\operatorname{Diff}(\mathcal{F}) \doteq \{\operatorname{Diff}(f) : f \in \mathcal{F}\}$.

Given any $a, b \in \mathbb{R}$, we use $a \wedge b \doteq \min(a, b)$ and $a \vee b \doteq \max(a, b)$. We use the Iverson bracket (Knuth, 1992) $[\![p]\!]$ to denote the indicator function, whose value is 1 if p is true and 0 otherwise. For any $x_0 \in \mathbb{R}$, we use $\delta_{x_0}(\cdot)$ to denote the Dirac delta centred at x_0, which is a generalised function[2] satisfying $\int_{\mathbb{R}} \delta_{x_0}(x) f(x)\, dx = f(x_0)$ for any continuous $f : \mathbb{R} \to \mathbb{R}$.

For any $z \in \mathbb{R}$, we define $\operatorname{sign}(z) = [\![z \geq 0]\!] - [\![z \leq 0]\!]$. The sigmoid function $\sigma(\cdot)$ is defined by

$$(\forall z \in \mathbb{R})\, \sigma(z) \doteq \frac{1}{1 + e^{-z}}, \tag{2}$$

with its inverse $\sigma^{-1}(\cdot)$ being the logit function,

$$(\forall y \in (0, 1))\, \sigma^{-1}(y) \doteq \log \frac{y}{1 - y}.$$

2.2 Probability distributions and random variables

We use sans-serif fonts e.g. X, Y to denote random variables. We denote by $\mathsf{X} \sim D$ that X is a random variable with probability distribution D. We denote by $\mathbb{P}_{\mathsf{X}\sim D}[\mathsf{X} \in \mathcal{A}]$ the probability that a random draw of X according to D falls in the set $\mathcal{A}$. We denote by $\mathop{\mathbb{E}}_{\mathsf{X}\sim D}[\mathsf{X}]$ the expected value of the random variable X.

Given distributions P and Q such that P is absolutely continuous with respect to Q, we use $\frac{dP}{dQ}$ to denote the Radon-Nikodym density of P with respect to Q. When it exists, we refer to the density of a random variable with respect to Lebesgue measure (unless noted otherwise) by p_{X}. Alternately, when the random variable is clear from context, we refer to the density of the underlying distribution (e.g. Q) by the corresponding lowercase letter (e.g. q).

Given a set $\mathcal{S}$, we denote by $\Delta_{\mathcal{S}}$ by the set of all probability distributions on $\mathcal{S}$. We denote by $\operatorname{Ber}(\theta)$ the Bernoulli distribution with parameter $\theta \in [0, 1]$.

2.3 f- and Bregman-divergences

For convex $f : \mathbb{R}_+ \to \mathbb{R}$, the f-divergence (Csiszár, 1963) between distributions P, Q is

$$\mathbb{I}_f(P, Q) \doteq \mathop{\mathbb{E}}_{\mathsf{X}\sim Q}\left[\left(\frac{dP}{dQ}\right)(\mathsf{X})\right]. \tag{3}$$

This can be seen as a notion of discrepancy between P and Q. For normalisation purposes, one typically enforces $f(1) = 0$.

A *generative Bregman divergence* is a distinct notion of discrepancy between two probability distributions. It relies on the notion of a *Bregman divergence* (Bregman, 1967). For convex, differentiable $f : \mathbb{R} \to \mathbb{R}$, the Bregman divergence B_f between points $x, y \in \mathbb{R}$ is

$$B_f(x, y) \doteq f(x) - f(y) - f'(y) \cdot (x - y).$$

The generative Bregman divergence $\mathbb{B}_f$ between distributions P, Q with densities p, q with respect to some distribution M is then the average divergence between the densities (Reid and Williamson, 2011, Section 3.3),

$$\mathbb{B}_f(P, Q) \doteq \mathop{\mathbb{E}}_{\mathsf{X}\sim M}[B(p(\mathsf{X}), q(\mathsf{X}))]. \tag{4}$$

2. Strictly, the Dirac delta is defined as a distribution or functional such that $\delta_{x_0}(f) = f(x_0)$ for every smooth f (Rudin, 1973, pg. 156), (Strichartz, 1994, pg. 5), or one interprets the integral $\int_{\mathbb{R}} \delta_{x_0}(x) f(x)\, dx$ as $\int_{\mathbb{R}} f(x)\, \mu_{x_0}(dx)$ for μ_{x_0} being the Dirac measure.

2.4 Scorers and pair-scorers

We are interested in supervised learning problems involving an instance or feature space $\mathcal{X}$ (often $\mathbb{R}^n$), and a label space $\mathcal{Y}$. We call an element $x \in \mathcal{X}$ an *instance* or *feature vector*, and an element $y \in \mathcal{Y}$ a *label*. A *scorer* s for the space $(\mathcal{X}, \mathcal{Y})$ is some (measurable) function $s : \mathcal{X} \to \mathcal{V}$, where $\mathcal{V} \subseteq \mathbb{R}^{|\mathcal{Y}|}$ is the prediction space of the scorer. The magnitude of each element of s corresponds to the degree of belief in an instance having the corresponding label. A *classifier* is a scorer with $\mathcal{V} = \mathcal{Y}$, so that an instance is directly annotated with one of the labels. A *class-probability estimator* is a scorer with $\mathcal{V} = \Delta_{[|\mathcal{Y}|]}$, so that an instance is annotated by a distribution over its possible labels.

This paper focusses on the setting of *binary labels*, where $\mathcal{Y} = \{\pm 1\}$. In the case of binary labels, a scorer is some $s : \mathcal{X} \to \mathcal{V}$, where $\mathcal{V} \subseteq \mathbb{R}$. Here, classifier c is often derived from a scorer[3] s via $c(x; t) = 2 [\![s(x) \geq t]\!] - 1$ for some threshold $t \in \mathbb{R}$. Similarly, a class-probability estimator f is often derived from a scorer s via $f = \Psi^{-1} \circ s$ for some (inverse) link function $\Psi^{-1} : \mathbb{R} \to [0, 1]$.

When a scorer is applied to instances drawn from some distribution, one can consider the induced distribution over scores. If X is a random variable over instances with distribution $\boldsymbol{M}$, we denote by $\boldsymbol{S}$ the induced distribution over the scores.When it exists, we refer to the induced (marginal) distribution of the scorer as $\boldsymbol{M}_S$, and the distributions on the positive and negative classes by $\boldsymbol{P}_S$ and $\boldsymbol{Q}_S$ respectively. We denote the marginal density of the score distribution by μ_S, and the score densities on the positive and negative classes by p_S and q_S respectively.

A *pair-scorer* s_{Pair} for a product space $\mathcal{X} \times \mathcal{X}$ is some function $s_{\text{Pair}} : \mathcal{X} \times \mathcal{X} \to \mathcal{V}$. The magnitude of s_{Pair} corresponds to a degree of belief in the first instance having a "larger" label than the second, according to some metric. A *ranker* is a pair-scorer with $\mathcal{V} = \{\pm 1\}$. A ranker r is typically derived from a pair-scorer s_{Pair} via $r(x, x'; t) = 2 [\![s(x, x') \geq t]\!] - 1$ for some threshold $t \in \mathbb{R}$. Given a (standard, or univariate) scorer s, we can construct a pair-scorer $\text{Diff}(s) : \mathcal{X} \times \mathcal{X} \to \mathcal{V} - \mathcal{V}$ (where $-$ denotes Minkowski subtraction) via

$$(\forall x, x' \in \mathcal{X}) \, \text{Diff}(s)(x, x') \doteq s(x) - s(x').$$

We call a pair-scorer s_{Pair} *decomposable* if

$$s_{\text{Pair}} \in \mathcal{S}_{\text{Decomp}} \doteq \{ s_{\text{Pair}} \in \mathcal{V}^{\mathcal{X} \times \mathcal{X}} : (\exists s : \mathcal{X} \to \mathbb{R}) \, s_{\text{Pair}} = \text{Diff}(s) \}.$$

We call a pair-scorer *anti-symmetric* if, for every $x, x' \in \mathcal{X}$, $s_{\text{Pair}}(x, x') = -s_{\text{Pair}}(x', x)$. Every decomposable scorer is anti-symmetric, but not conversely.

2.5 Calibration transform

Given a scorer s and distribution $D \in \Delta_{\mathcal{X} \times \{\pm 1\}}$, the *score-to-probability transform* $\text{Prb}(\cdot; D, s) : \mathbb{R} \to [0, 1]$ maps each score to the actual probability when the score is observed:

$$(\forall a \in \text{Im}(s)) \, \text{Prb}(a; D, s) \doteq \mathbb{P}_{(\mathsf{X}, \mathsf{Y}) \sim D}[\mathsf{Y} = 1 | s(\mathsf{X}) = a].$$

We call a scorer s *calibrated* with respect to D if each predicted score equals the probability of $\mathsf{Y} = 1$ when that prediction is made (DeGroot and Fienberg, 1983):

$$(\forall a \in \text{Im}(s)) \, \text{Prb}(a; D, s) = a.$$

A scorer must be a class-probability estimator to be calibrated (i.e. it cannot output values outside $[0, 1]$). The *calibration transform* $\text{Cal}(\cdot; D, s) : \mathcal{X} \to [0, 1]$ converts a scorer into a class-probability estimator via

$$(\forall x \in \mathcal{X}) \, \text{Cal}(x; D, s) \doteq \text{Prb}(s(x); D, s). \tag{5}$$

It is easy to check that for any scorer $s : \mathcal{X} \to \mathbb{R}$, $\text{Cal}(\cdot\,; D, s)$ is automatically calibrated.

3. The case of $s(x) = t$ can be considered as a tie, thus requiring a tie-breaking scheme. The above definition corresponds to breaking ties in favour of the positive class.

2.6 Loss functions and conditional risks

A *binary classification loss* ℓ, often just referred to as a *loss*, is some (measurable) function $\ell : \{\pm 1\} \times \mathbb{R} \to \mathbb{R}_+$. An important example is the 0-1 loss[4],

$$\ell_{01}(y, v) \doteq [\![yv < 0]\!] + \frac{1}{2} \cdot [\![v = 0]\!] .$$

Given a loss ℓ, we use $\ell_1(v) = \ell(1, v)$ and $\ell_{-1}(v) = \ell(-1, v)$ to denote the individual *partial losses*. We will sometimes refer to a loss via the tuple $\ell(v) = (\ell_{-1}(v), \ell_1(v))$.

We define the *conditional ℓ-risk* $L(\cdot, \cdot; \ell) \colon [0, 1] \times \mathbb{R} \to \mathbb{R}_+$ to be

$$(\forall \eta \in [0, 1], s \in \mathbb{R})\ L(\eta, s; \ell) \doteq \mathop{\mathbb{E}}_{\mathsf{Y} \sim \mathrm{Ber}(\eta)} [\ell(\mathsf{Y}, s)] = \eta \cdot \ell_1(s) + (1 - \eta) \cdot \ell_{-1}(s). \tag{6}$$

The *Bayes-optimal conditional ℓ-risk* $L^*(\cdot; \ell) \colon [0, 1] \to \mathbb{R}_+$ is then the best possible conditional risk,

$$(\forall \eta \in [0, 1])\ L^*(\eta; \ell) \doteq \inf_{s \in \mathbb{R}} L(\eta, s; \ell).$$

For the 0-1 loss, the optimal risk is attained for any score with the same sign as $\eta - \frac{1}{2}$. More generally, we call a loss *classification calibrated* if for every $\eta \in [0, 1] \setminus \left\{ \frac{1}{2} \right\}$ (Bartlett et al., 2006, Definition 1),

$$L^*(\eta; \ell) < \inf_{s \colon s \cdot (2\eta - 1) \leq 0} L(\eta, s; \ell), \tag{7}$$

i.e. every optimal prediction has the same sign as $\eta - \frac{1}{2}$.

We call a loss ℓ *symmetric* if, for every $y \in \{\pm 1\}$ and $v \in \mathbb{R}$, $\ell(y, v) = \ell(-y, -v)$. We denote the *symmetrised version* of an arbitrary loss by

$$\ell_{\mathrm{symm}}(v) \doteq \frac{\ell(1, v) + \ell(-1, -v)}{2}. \tag{8}$$

We call a loss ℓ a *margin loss* if $\ell(y, z) = \phi(yz)$ for some function $\phi : \mathbb{R} \to \mathbb{R}_+$. A loss is symmetric if and only if it is a margin loss; sufficiency is straightforward, and to see necessity, note that for a symmetric loss, $\ell(y, v) = [\![y = 1]\!]\, \ell_1(v) + [\![y = -1]\!]\, \ell_1(-v) = \ell_1(yv) = \phi(yv)$ for $\phi(v) \doteq \ell_1(v)$.

2.7 Proper and proper-composite losses

A *probability estimation loss* λ is some (measurable) function $\lambda : \{\pm 1\} \times [0, 1] \to \mathbb{R}_+ \cup \{+\infty\}$. We call a probability estimation loss *proper*[5] if its conditional risk is optimised by predicting the underlying probability (Buja et al., 2005; Reid and Williamson, 2010),

$$(\forall \eta, \eta' \in [0, 1])\ L(\eta, \eta; \lambda) \leq L(\eta, \eta'; \lambda). \tag{9}$$

We call a loss *strictly proper* if the inequality is strict.

In the following, we assume two mild regularity conditions: that $\lambda_1(1) = \lambda_{-1}(0) = 0$, and

$$\lim_{u \to 0} u \cdot \lambda_1(u) = \lim_{u \to 1} (1 - u) \cdot \lambda_{-1}(u) = 0.$$

4. When $\mathcal{V} = \{\pm 1\}$, so that we are assessing a classifier, the canonical definition of 0-1 loss is $\ell_{01}(y, v) = [\![y \neq v]\!]$. When assessing a scorer with $\mathcal{V} = \mathbb{R}$, we simply derive a classifier from the scores and compute the resulting 0-1 loss. The only mild complication is that we consider a score of $v = 0$ to be a tie, which thus requires a tie-breaking scheme. In our definition, we break ties uniformly at random, and thus generate a *randomised* classifier. This results in the extra second term compared to the classification case.
5. Proper losses are sometimes referred to as proper *scoring rules* (Gneiting and Raftery, 2007), especially in the statistics literature. A "scoring rule" is distinct from our notion of a "scorer": the former is a loss, and the latter is a prediction. In the literature on scoring rules, our notion of a scorer is sometimes referred to as a (probabilistic) "forecast" (Gneiting and Katzfuss, 2014).

Name	$\lambda_{-1}(u)$	$\lambda_1(u)$	$w(c)$	$L^*(\eta;\lambda)$
0-1	$\left[\!\left[u > \frac{1}{2}\right]\!\right]$	$\left[\!\left[u < \frac{1}{2}\right]\!\right]$	$\delta_{1/2}(c)$	$\eta \wedge (1-\eta)$
Cost-sensitive	$c^* \cdot \left[\!\left[u > \frac{1}{2}\right]\!\right]$	$(1-c^*) \cdot \left[\!\left[u < \frac{1}{2}\right]\!\right]$	$\delta_{c^*}(c)$	$((1-c^*)\cdot\eta) \wedge (c^* \cdot (1-\eta))$
Brier	u^2	$(1-u)^2$	2	$\eta \cdot (1-\eta)$
Log	$-\log(1-u)$	$-\log u$	$\frac{1}{c\cdot(1-c)}$	$-\eta\cdot\log\eta - (1-\eta)\cdot\log(1-\eta)$
Boosting	$\left(\frac{u}{1-u}\right)^{1/2}$	$\left(\frac{1-u}{u}\right)^{1/2}$	$\frac{1}{2\cdot(c\cdot(1-c))^{3/2}}$	$2\cdot\sqrt{\eta\cdot(1-\eta)}$

Table 3: Examples of proper losses λ with associated weight functions w and conditional Bayes risks $L^*(\cdot;\lambda)$.

Name	Symbol	$\ell(y,v)$	λ	$\Psi(u)$	$\Psi^{-1}(v)$
Square	ℓ_{sq}	$\frac{1}{4}\cdot(1-yv)^2$	Brier	$2u-1$	$\left(\frac{v+1}{2} \vee 0\right) \wedge 1$
Logistic	$\ell_{\log}$	$\log(1+e^{-yv})$	Log	$\log\frac{u}{1-u}$	$\frac{1}{1+e^{-v}}$
Exponential	$\ell_{\exp}$	e^{-yv}	Boosting	$\frac{1}{2}\cdot\log\frac{u}{1-u}$	$\frac{1}{1+e^{-2v}}$
Matsushita	ℓ_{mts}	$\sqrt{1+\frac{v^2}{4}} - \frac{yv}{2}$	Boosting	$\frac{2u-1}{\sqrt{u\cdot(1-u)}}$	$\frac{1}{2}\cdot\left(1+\frac{v/2}{\sqrt{1+(v/2)^2}}\right)$

Table 4: Examples of proper composite losses with associated underlying proper loss λ and link function Ψ.

Any proper loss satisfying these conditions admits an integral representation as a weighted combination of cost-sensitive losses (Shuford Jr. et al., 1966), (Schervish, 1989, Theorem 4.2),

$$\lambda : (y,u) \mapsto \int_0^1 w(c)\cdot\lambda_{\text{CS}(c)}(y,u)\,dc, \tag{10}$$

where $w : [0,1] \to \mathbb{R}_+ \cup \{+\infty\}$ is called the *weight function* of the loss, and $\lambda_{\text{CS}(c)}$ is the *cost-sensitive loss*

$$\begin{aligned}\lambda_{\text{CS}(c)}(+1,u) &\doteq (1-c)\cdot[\![u<c]\!] + \frac{1}{2}\cdot[\![u=c]\!]\\ \lambda_{\text{CS}(c)}(-1,u) &\doteq c\cdot[\![u>c]\!] + \frac{1}{2}\cdot[\![u=c]\!].\end{aligned} \tag{11}$$

We call Equation 10 *Shuford's representation*. A loss is strictly proper if and only if its weight function is strictly positive. One can relate the weight function and conditional Bayes risk via (Reid and Williamson, 2010, Corollary 3) $w(c) = -(L^*(\cdot;\lambda))''(c)$.

We call a loss ℓ (*strictly*) *proper composite* if there is some invertible *link function* $\Psi : [0,1] \to \mathbb{R}\cup\{\pm\infty\}$ such that the probability estimation loss $\lambda(y,u) \doteq \ell(y,\Psi(u))$ is (strictly) proper (Reid and Williamson, 2010). By Equation 10, this implies that a proper composite loss also admits an integral representation; hence, we define the weight function of a proper composite loss as that of its underlying proper loss. We denote the set of strictly proper composite losses with invertible link function Ψ by $\mathcal{L}_{\text{SPC}}(\Psi)$, and the set of all proper composite losses by $\mathcal{L}_{\text{SPC}}$.

When the proper composite loss ℓ is differentiable, we have (Reid and Williamson, 2010, Corollary 12)

$$(\forall v \in \mathbb{R})\,\Psi^{-1}(v) = \left(1 - \frac{\ell_1'(v)}{\ell_{-1}'(v)}\right)^{-1}. \tag{12}$$

Given a proper loss λ, we call a link function $\Psi : [0,1] \to \mathbb{R}\cup\{\pm\infty\}$ *canonical* for that loss if

$$(\forall c \in [0,1])\,\Psi'(c) = w(c). \tag{13}$$

Problem	Input space	Output	Risk
Classification	$\mathcal{X} \times \mathcal{Y}$	$c : \mathcal{X} \to \mathcal{Y}$	$\mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim D}[\ell(\mathsf{Y}, s(\mathsf{X}))]$
Class-probability estimation	$\mathcal{X} \times \mathcal{Y}$	$\hat{\eta} : \mathcal{X} \to \Delta_{[\lvert\mathcal{Y}\rvert]}$	
Bipartite ranking	$\mathcal{X} \times \{\pm 1\}$	$s : \mathcal{X} \to \mathbb{R}$	$\mathop{\mathbb{E}}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q} \ell_{\text{symm}}(s(\mathsf{X}) - s(\mathsf{X}'))$
Pairwise ranking	$\mathcal{X} \times \mathcal{X} \times \{\pm 1\}$	$s_{\text{Pair}} : \mathcal{X} \times \mathcal{X} \to \{\pm 1\}$	$\mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{X}',\mathsf{Z})\sim R}\left[\ell(\mathsf{Z}, s_{\text{Pair}}(\mathsf{X}, \mathsf{X}'))\right]$

Table 5: Summary of learning problems in terms of input, output, and statistical risk.

The canonical link function is monotone increasing, but is strictly so if and only if the loss λ is strictly proper. For a strictly proper loss λ, the proper composite loss $\ell(y, v) = \lambda(y, \Psi^{-1}(v))$ is convex (Reid and Williamson, 2010, Theorem 28).

Table 3 provides some examples of popular proper losses, and Table 4 of popular proper composite losses, along with their associated underlying proper loss λ and link function Ψ.

3. Classification and ranking: statistical setups

We now formally define the problems of interest in this paper. For each problem, we state the nature of their assumed input, produced output, and measure of statistical performance (or *risk*). Table 5 provides a summary of these problems. Our starting point is a general statistical perspective on learning from binary labels.

3.1 Learning from binary labels: distributions and their decompositions

Most problems of interest in this paper concern learning based on samples from some distribution $D \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ over binary labels. The precise nature of these problems shall be specified momentarily, but we first note two decompositions of D that shall prove useful. The first involves splitting any $D \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ into

$$\begin{aligned}(\forall \mathcal{A} \subseteq \mathcal{X})\, P(\mathcal{A}) &= \mathbb{P}[\mathsf{X} \in \mathcal{A} | \mathsf{Y} = 1] \\ (\forall \mathcal{A} \subseteq \mathcal{X})\, Q(\mathcal{A}) &= \mathbb{P}[\mathsf{X} \in \mathcal{A} | \mathsf{Y} = -1] \\ \pi &= \mathbb{P}[\mathsf{X} \in \mathcal{X}, \mathsf{Y} = 1].\end{aligned}$$

Equivalently, we may decompose $D \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ into

$$\begin{aligned}(\forall \mathcal{A} \subseteq \mathcal{X})\, M(\mathcal{A}) &= \mathbb{P}[\mathsf{X} \in \mathcal{A}, \mathsf{Y} \in \{\pm 1\}] \\ (\forall x \in \mathcal{X})\, \eta(x) &= \mathbb{P}[\mathsf{Y} = 1 | \mathsf{X} = x].\end{aligned}$$

We refer to P, Q as the *class conditional distributions*, and π the *base rate*; we refer to M as the *observation distribution* and η as the *observation-conditional distribution* or *class-probability function*. We will denote the densities of P, Q with respect to M by p, q. When M possesses a density, we refer to it as μ. When we wish to refer to these constitutent distributions of D and their densities, we will explicitly parameterise D as either $D = \langle P, Q, \pi \rangle$ or $D = \langle M, \eta \rangle$ as appropriate.

We now proceed to formalising the problems considered in this paper.

3.2 Classification and class-probability estimation

Classification is the canonical supervised machine learning task, and has received several decades' worth of study from a theoretical and practical perspective. Classification is often motivated by appealing to several

real-world problems, such as predicting whether an email message is spam based on its contents, predicting whether or not a person is sick based on test results, or predicting whether or not an individual will enjoy a movie based on its characteristics.

Formally, in statistical classification (Devroye et al., 1996), we are given samples from some distribution $D \in \Delta_{\mathcal{X} \times \mathcal{Y}}$, and wish to learn a classifier $c : \mathcal{X} \to \mathcal{Y}$. In *class-probability estimation* (Buja et al., 2005; Reid and Williamson, 2010), the input is identical, except we wish to learn a class-probability estimator $\hat{\eta} : \mathcal{X} \to \Delta_{[|\mathcal{Y}|]}$. In the binary case, classification involves learning some $c : \mathcal{X} \to \{\pm 1\}$ and class-probability estimation involves learning some $\hat{\eta} : \mathcal{X} \to [0, 1]$. In this setting, one typically first learns a general scorer $s : \mathcal{X} \to \mathbb{R}$, and performs either thresholding to get a classifier or a monotone transformation to get a class-probability estimator.

It remains to define how the performance of a candidate scorer $s : \mathcal{X} \to \mathbb{R}$ is assessed for the two problems. Intuitively, as these problems assume examples drawn from a probability distribution, one would like to incur a small *disutility* or *loss* for a *randomly drawn* example. This notion is captured by the notion of *statistical risk*. Given any $D = \langle M, \eta \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$ and loss ℓ, we define the *ℓ-classification risk* for a scorer s to be the average loss incurred on a random sample from D,

$$\mathbb{L}(s; D, \ell) \doteq \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y}) \sim D} [\ell(\mathsf{Y}, s(\mathsf{X}))] = \mathop{\mathbb{E}}_{\mathsf{X} \sim M} [L(\eta(\mathsf{X}), s(\mathsf{X}); \ell)], \tag{14}$$

recalling that $L(\cdot, \cdot; \ell)$ is the conditional ℓ-risk (Equation 6). The *Bayes-optimal ℓ-classification risk*, or simply the *Bayes risk*, is the infimal ℓ-classification risk:

$$\mathbb{L}^*(D, \ell) \doteq \inf_{s \,:\, \mathcal{X} \to \mathbb{R}} \mathbb{L}(s; D, \ell).$$

Similarly, the *Bayes-optimal conditional risk*, or simply the *conditional Bayes risk* is

$$L^*(\eta; \ell) \doteq \inf_{s \in \mathbb{R}} L(\eta, s; \ell).$$

When the infimum is achievable[6], the set of *Bayes-optimal scorers* for a loss ℓ and distribution D comprises scorers that attain the Bayes-optimal ℓ-classification risk:

$$\mathcal{S}^*(D, \ell) \doteq \operatorname*{Argmin}_{s \,:\, \mathcal{X} \to \mathbb{R}} \mathbb{L}(s; D, \ell).$$

Under appropriate measurability assumptions, this set may be discerned pointwise, by studying the minimisers of the conditional risk $L(\cdot, \cdot; \ell)$ (Steinwart, 2007). Finally, the *ℓ-regret* of a scorer $s : \mathcal{X} \to \mathbb{R}$ is its excess ℓ-classification risk over the Bayes ℓ-classification risk:

$$\operatorname{regret}(s; D, \ell) \doteq \mathbb{L}(s; D, \ell) - \mathbb{L}^*(D, \ell).$$

In binary classification, the canonical goal is to minimise the risk for ℓ_{01},

$$\mathbb{L}(s; D, \ell_{01}) = \mathbb{P}_{(\mathsf{X},\mathsf{Y}) \sim D} [\mathsf{Y} \cdot s(\mathsf{X}) < 0] + \frac{1}{2} \cdot \mathbb{P}_{(\mathsf{X},\mathsf{Y}) \sim D}[s(\mathsf{X}) = 0], \tag{15}$$

i.e. we want our scorer s to achieve low misclassification error on future samples. Observe that the second term above encodes that ties in the classification are broken uniformly at random. In class probability estimation on the other hand, the canonical goal is to minimise $\mathbb{L}(s; D, \ell)$ for some proper composite loss ℓ.

The binary classification ℓ-risk has been extensively studied. Much of the literature has focussed on the design of losses ℓ with favourable computational and statistical properties, with specific focus on margin losses such as in the support vector machine (which employs the hinge loss $\ell_{\text{hinge}}(y, v) = \max(0, 1 - yv)$), boosting (which employs the exponential loss $\ell_{\exp}(y, v) = e^{-yv}$), and logistic regression (which employs the logistic loss $\ell_{\log}(y, v) = \log(1 + e^{-yv})$). There has also been theoretical study of how regret with respect to a loss ℓ translates into the regret with respect to the 0-1 loss, via *surrogate regret bounds* (Zhang, 2004; Bartlett et al., 2006).

6. A simple example where this is not true is the case of separable data, where $\eta(x) \in \{0, 1\}$ for every x. Under log-loss, which is proper, the optimal scorer $s^*(x) = \eta(x)$. Under logistic loss, which is proper composite, the optimal scorer is in general $s^*(x) = \log \frac{\eta(x)}{1-\eta(x)}$. But for separable data, that would require $s^*(x) \in \{\pm\infty\}$, and so the infimum is not attainable. Working with the extended reals $\mathbb{R} \cup \{\pm\infty\}$ is one possible fix, but in the sequel we will always assume the infimum is attainable.

3.3 Instance ranking

Many applications traditionally used to motivate binary classification are more appropriately cast as *ranking* problems. For example, rather than merely predicting an individual's enjoyment of a single movie, it is potentially more useful to have a system capable of producing a *ranked list* of movies that the individual may enjoy. Similarly, in epidemiological studies, rather than merely predicting whether a single individual has a disease, it is potentially more useful to have a system that can produced a ranked list of individuals deemed most likely to have a particular disease. Instance ranking is of similar interest in most other settings where classification is, such as credit fraud detection and clickthrough rate analysis.

Formally, in *instance ranking* (Fürnkranz and Hüllermeier, 2010, pg. 6) (Crammer and Singer, 2001; Shashua and Levin, 2002), we are given samples from some distribution $D \in \Delta_{\mathcal{X}\times\mathcal{Y}}$, where each sample comprises a single $x \in \mathcal{X}$ and a label $y \in \mathcal{Y}$, where $\mathcal{Y}$ is a (finite) totally ordered set. The goal is to learn a scorer $s : \mathcal{X} \to \mathbb{R}$ such that the ordering of instances by their scores scores mimics the ordering by their labels. Note that $\mathcal{Y}$, while ordered, does not necessarily have an associated metric e.g. for $\mathcal{Y} = \{\text{Hate}, \text{Indifferent}, \text{Like}\}$, there is a natural ordering over the outcomes, but it may not be possible to assign numeric values to the comparison of two outcomes. When $\mathcal{Y}$ does not have an associated metric, instance ranking is identical to *ordinal regression* (Agresti, 1984), and thus one can solve ordinal regression problems by ranking methods (Herbrich et al., 2000). When $\mathcal{Y}$ has an associated metric, instance ranking is a hybrid between traditional multi-class classification (where $\mathcal{Y}$ is finite but unordered) and regression (where $\mathcal{Y}$ is ordered but not finite) (Li and Lin, 2006).

When $\mathcal{Y} = \{\pm 1\}$, the goal of instance ranking can be seen as scoring positive examples higher than negative examples. This special case is known as the *bipartite ranking* problem, and has received considerable attention (Freund et al., 2003; Agarwal and Niyogi, 2005; Clémençon et al., 2008; Kotlowski et al., 2011). Bipartite ranking is the main focus of this paper.

As per the previous section, to specify the bipartite ranking problem, we must begin with its underlying risk. Given any $D = \langle P, Q, \pi \rangle \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ and loss ℓ, we define the *ℓ-bipartite risk* for a scorer $s : \mathcal{X} \to \mathbb{R}$ to be

$$\begin{aligned}\mathbb{L}_{\text{BR}}(s; D, \ell) &\doteq \mathop{\mathbb{E}}_{\mathsf{X}\sim P, \mathsf{X}'\sim Q} \left[\ell_{\text{symm}}(s(\mathsf{X}) - s(\mathsf{X}'))\right] \\ &= \frac{1}{\pi \cdot (1-\pi)} \mathop{\mathbb{E}}_{\mathsf{X}\sim M, \mathsf{X}'\sim M} \left[\eta(\mathsf{X}) \cdot (1 - \eta(\mathsf{X}')) \cdot \ell_{\text{symm}}(s(\mathsf{X}) - s(\mathsf{X}'))\right]. \end{aligned} \tag{16}$$

The first equation indicates that this risk is *independent* of the base rate π. When the loss ℓ is symmetric, the equation reduces to

$$\mathbb{L}_{\text{BR}}(s; D, \ell) = \mathop{\mathbb{E}}_{\mathsf{X}\sim P, \mathsf{X}'\sim Q} \left[\ell_1(s(\mathsf{X}) - s(\mathsf{X}'))\right]. \tag{17}$$

A canonical goal is to minimise the bipartite risk for ℓ_{01}, which is

$$\mathbb{L}_{\text{BR}}(s; D, \ell_{01}) = \mathbb{P}_{\mathsf{X}\sim P, \mathsf{X}'\sim Q}[s(\mathsf{X}) - s(\mathsf{X}') < 0] + \frac{1}{2} \cdot \mathbb{P}_{\mathsf{X}\sim P, \mathsf{X}'\sim Q}[s(\mathsf{X}) - s(\mathsf{X}') = 0],$$

i.e. we want s to achieve low pairwise misclassification error, in the sense of scoring a negative higher than a positive, on future samples. The reader may wonder how, if at all, this relates to the misclassification error of Equation 15; such discussion shall be deferred to §10.

Equipped with a notion of statistical risk, we can define the Bayes-optimal risk, Bayes-optimal scorers, and regret by analogy with the quantities from the previous section:

$$\begin{aligned}\mathbb{L}^*_{\text{BR}}(D, \ell) &\doteq \inf_{s:\ \mathcal{X}\to\mathbb{R}} \mathbb{L}_{\text{BR}}(s; D, \ell) \\ \mathcal{S}^*_{\text{BR}}(D, \ell) &\doteq \operatorname*{Argmin}_{s:\ \mathcal{X}\to\mathbb{R}} \mathbb{L}_{\text{BR}}(s; D, \ell) \\ \text{regret}_{\text{BR}}(s; D, \ell) &\doteq \mathbb{L}_{\text{BR}}(s; D, \ell) - \mathbb{L}^*_{\text{BR}}(D, \ell). \end{aligned} \tag{18}$$

As in binary classification, a large body of literature has focussed on the design of losses ℓ with favourable numerical and statistical properties, again with specific focus on margin losses such as in SVMRank (Joachims, 2002; Herbrich et al., 1998) (corresponding to hinge loss), RankBoost (Freund et al., 2003) (corresponding to exponential loss), and RankNet (Burges et al., 2005) (corresponding to logistic loss).

3.4 Pairwise ranking

Pairwise ranking problems involve labelled *pairs* of instances, where we only know which of two instances is more likely to possess some characteristic. For example, in web search click-log data, we can elicit pairwise preferences to determine which of two search results is more likely to be clicked (Joachims, 2002). (In information retrieval, the problem is sometimes referred to as "average view paper ranking" (Liu, 2009, pg. 203).)

Formally, in *pairwise ranking* (Cohen et al., 1999; Herbrich et al., 1998) (Fürnkranz and Hüllermeier, 2010, pg. 7), we are given samples from some distribution $R \in \Delta_{\mathcal{X}\times\mathcal{X}\times\{\pm 1\}}$, where each sample comprises pairs of instances $(x^{(1)}, x^{(2)})$, and a label $\{\pm 1\}$, denoting that $x^{(1)}$ ranks above or below $x^{(2)}$ respectively. Following the notation used for binary classification, we will write R as either $R = \langle P_{\text{pair}}, Q_{\text{pair}}, \pi_{\text{pair}}\rangle \in \Delta_{\mathcal{X}\times\mathcal{X}\times\{\pm 1\}}$ or $R = \langle M_{\text{pair}}, \eta_{\text{Pair}}\rangle$, where for example $(\forall \mathcal{A} \subseteq \mathcal{X}\times\mathcal{X})\, P_{\text{pair}}(\mathcal{A}) = \mathbb{P}[(\mathsf{X}, \mathsf{X}') \in \mathcal{A} | \mathsf{Z} = 1]$ for random variables $\mathsf{X}, \mathsf{X}', \mathsf{Z}$ defined over the instances and label respectively.

The goal in pairwise ranking is to learn a pair-classifier $c_{\text{Pair}} : \mathcal{X}\times\mathcal{X} \to \{\pm 1\}$ that specifies whether or not the first instance ranks above the second. As with standard classification, this is often done by thresholding a pair-scorer $s_{\text{Pair}} : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$. A pair-scorer is nothing more[7] than a scorer defined on $\mathcal{X}\times\mathcal{X}$, and thus the notion of risk for pairwise ranking is as expected: given any $R = \langle M_{\text{pair}}, \eta_{\text{Pair}}\rangle \in \Delta_{\mathcal{X}\times\mathcal{X}\times\{\pm 1\}}$ and loss ℓ, we define the *ℓ-pairwise ranking risk* for a pair-scorer s_{Pair} to be

$$\mathbb{L}(s_{\text{Pair}}; R, \ell) \doteq \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{X}',\mathsf{Z})\sim R} \left[\ell(\mathsf{Z}, s_{\text{Pair}}(\mathsf{X}, \mathsf{X}'))\right] = \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{X}')\sim M_{\text{pair}}} \left[L(\eta_{\text{Pair}}(\mathsf{X}, \mathsf{X}'), s_{\text{Pair}}(\mathsf{X}, \mathsf{X}'); \ell)\right]. \tag{19}$$

A canonical goal is to minimise this risk for the 0-1 loss,

$$\mathbb{L}(s_{\text{Pair}}; R, \ell_{01}) = \mathbb{P}_{(\mathsf{X},\mathsf{X}',\mathsf{Z})\sim R}[\mathsf{Z}\cdot s_{\text{Pair}}(\mathsf{X}, \mathsf{X}') < 0] + \frac{1}{2}\mathbb{P}_{(\mathsf{X},\mathsf{X}',\mathsf{Z})\sim R}[s_{\text{Pair}}(\mathsf{X}, \mathsf{X}') = 0],$$

i.e. we want our pair-scorer s_{Pair} to achieve low pairwise misclassification error on future samples. Observe that the second term above encodes that ties in the ranking are broken uniformly at random.

As in the previous sections, we may define the Bayes-optimal risk, Bayes-optimal scorers, and regret for pairwise ranking as:

$$\begin{aligned}
\mathbb{L}^*(R, \ell) &\doteq \inf_{s_{\text{Pair}} : \mathcal{X}\times\mathcal{X}\to\mathbb{R}} \mathbb{L}(s_{\text{Pair}}; R, \ell) \\
\mathcal{S}^*(R, \ell) &\doteq \operatorname*{Argmin}_{s_{\text{Pair}} : \mathcal{X}\times\mathcal{X}\to\mathbb{R}} \mathbb{L}(s_{\text{Pair}}; R, \ell) \\
\text{regret}(s_{\text{Pair}}; R, \ell) &\doteq \mathbb{L}(s_{\text{Pair}}; R, \ell) - \mathbb{L}^*(R, \ell).
\end{aligned}$$

4. Reducing bipartite to pairwise ranking

Having defined the risks for three seemingly disparate problems, it is worth noting how they relate to each other. First, the pairwise ranking risk (Equation 19) is clearly equivalent to the classification risk (Equation 14), except that we operate over pairs of instances $\mathcal{X}\times\mathcal{X}$. Thus, pairwise ranking is equivalent to classification on pairs; see also §10.1. Further, we may view bipartite ranking as a special case of pairwise ranking in the following sense: any $D \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ may be converted to a pairwise ranking distribution, $D_{\text{BR}} \in \Delta_{\mathcal{X}\times\mathcal{X}\times\{\pm 1\}}$, such that the risks of the two problems are equivalent. Thus, the methods described above for pairwise ranking are equally applicable for bipartite ranking.

7. Nonetheless, making a distinction with a standard scorer shall be useful in our subsequent analysis.

Definition 1 (Derived pairwise ranking distribution) *Given any* $D = \langle P, Q, \pi \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$, *the* derived pairwise ranking distribution $D_{\mathrm{BR}} \in \Delta_{\mathcal{X} \times \mathcal{X} \times \{\pm 1\}}$ *corresponds to the following process for drawing a* $(\mathsf{X}, \mathsf{X}', \mathsf{Z}) \in \mathcal{X} \times \mathcal{X} \times \{\pm 1\}$*:*

- *Draw* $\mathsf{Z} \sim \mathrm{Ber}(1/2)$
- *If* $\mathsf{Z} = +1$, *draw* $\mathsf{X} \sim P, \mathsf{X}' \sim Q$
- *If* $\mathsf{Z} = -1$, *draw* $\mathsf{X} \sim Q, \mathsf{X}' \sim P$.

An equivalent process is:

- *Draw* $(\mathsf{X}, \mathsf{Y}) \sim D$
- *Draw* $(\mathsf{X}', \mathsf{Y}') \sim D$.
- *If* $\mathsf{Y} = \mathsf{Y}'$, *reject and re-sample; else, let* $\mathsf{Z} = \mathsf{Y}$.

The following risk equivalence can be easily verified, and is well-known for the case of ℓ_{01} (Balcan et al., 2008; Kotlowski et al., 2011; Agarwal, 2014). (See Proposition 62 for a proof of a more general result).

Lemma 2 *For any distribution* $D = \langle P, Q, \pi \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$, *loss* ℓ *and scorer* $s : \mathcal{X} \to \mathbb{R}$,

$$\mathbb{L}_{\mathrm{BR}}(s; D, \ell) = \mathbb{L}(\mathrm{Diff}(s); D_{\mathrm{BR}}, \ell).$$

Proof By Equation 16,

$$\begin{aligned}
\mathbb{L}_{\mathrm{BR}}(s; D, \ell) &= \mathop{\mathbb{E}}_{\mathsf{X} \sim P, \mathsf{X}' \sim Q} \left[\ell_{\mathrm{symm}}(s(\mathsf{X}) - s(\mathsf{X}')) \right] \\
&= \frac{1}{2} \cdot \mathop{\mathbb{E}}_{\mathsf{X} \sim P, \mathsf{X}' \sim Q} \left[\ell_1(s(\mathsf{X}) - s(\mathsf{X}')) + \ell_{-1}(s(\mathsf{X}') - s(\mathsf{X})) \right] \\
&= \frac{1}{2} \cdot \mathop{\mathbb{E}}_{\mathsf{X} \sim P, \mathsf{X}' \sim Q} \left[\ell_1(s(\mathsf{X}) - s(\mathsf{X}')) \right] + \frac{1}{2} \cdot \mathop{\mathbb{E}}_{\mathsf{X} \sim P, \mathsf{X}' \sim Q} \left[\ell_{-1}(s(\mathsf{X}') - s(\mathsf{X})) \right] \\
&= \frac{1}{2} \cdot \mathop{\mathbb{E}}_{\mathsf{X} \sim P, \mathsf{X}' \sim Q} \left[\ell_1(s(\mathsf{X}) - s(\mathsf{X}')) \right] + \frac{1}{2} \cdot \mathop{\mathbb{E}}_{\mathsf{X} \sim Q, \mathsf{X}' \sim P} \left[\ell_{-1}(s(\mathsf{X}') - s(\mathsf{X})) \right] \\
&= \frac{1}{2} \cdot \mathop{\mathbb{E}}_{(\mathsf{X}, \mathsf{X}') \sim (P \times Q)} \left[\ell_1(s(\mathsf{X}) - s(\mathsf{X}')) \right] + \frac{1}{2} \cdot \mathop{\mathbb{E}}_{(\mathsf{X}, \mathsf{X}') \sim (Q \times P)} \left[\ell_{-1}(s(\mathsf{X}') - s(\mathsf{X})) \right],
\end{aligned}$$

where in the penultimate equation we have simply renamed the random variables in the second expression.

By definition of D_{BR} and the pairwise ranking risk (Equation 19), this is exactly $\mathbb{L}(\mathrm{Diff}(s); D_{\mathrm{BR}}, \ell)$. ■

We summarise the conditional and marginal probabilities of D_{BR} in Appendix B. For example, if $(\mathsf{X}, \mathsf{X}', \mathsf{Z})$ denotes the random variables distributed according to D_{BR}, it is not hard to check that

$$\begin{aligned}
\mathbb{P}[(\mathsf{X}, \mathsf{X}') \in \mathcal{A} \times \mathcal{B} | \mathsf{Z} = +1] &= P(\mathcal{A}) \cdot Q(\mathcal{B}) \\
\mathbb{P}[(\mathsf{X}, \mathsf{X}') \in \mathcal{A} \times \mathcal{B} | \mathsf{Z} = -1] &= P(\mathcal{B}) \cdot Q(\mathcal{A}).
\end{aligned}$$

The risk equivalence in Lemma 2 has a subtlety that will prove important in our subsequent analysis: the bipartite risk for a scorer relates to the pairwise risk of a *decomposable* pair-scorer. Consequently, we cannot in general equate the *Bayes-optimal* bipartite risks for the two problems, as the following makes precise.

Proposition 3 *For any distribution* $D \in \Delta_{\mathcal{X} \times \{\pm 1\}}$ *and loss* ℓ,

$$\mathbb{L}^*_{\mathrm{BR}}(D, \ell) = \mathbb{L}^*(D_{\mathrm{BR}}, \ell) \iff \mathcal{S}^*(D_{\mathrm{BR}}, \ell) \cap \mathcal{S}_{\mathrm{Decomp}} \neq \emptyset.$$

Proof By definition,

$$\begin{aligned}\mathbb{L}^*_{\mathrm{BR}}(D, \ell) &= \inf_{s\colon \mathcal{X}\to\mathbb{R}} \mathbb{L}_{\mathrm{BR}}(s; D, \ell) \\ &= \inf_{s\colon \mathcal{X}\to\mathbb{R}} \mathbb{L}(\mathrm{Diff}(s); D_{\mathrm{BR}}, \ell) \text{ by Lemma 2} \\ &= \inf_{s_{\mathrm{Pair}}\in\mathcal{S}_{\mathrm{Decomp}}} \mathbb{L}(s_{\mathrm{Pair}}; D_{\mathrm{BR}}, \ell).\end{aligned}$$

By contrast,

$$\mathbb{L}^*(D_{\mathrm{BR}}, \ell) = \inf_{s_{\mathrm{Pair}}\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}} \mathbb{L}(s_{\mathrm{Pair}}; D_{\mathrm{BR}}, \ell).$$

The two Bayes-risks involve minimisation of the same functional, $\mathbb{L}(s_{\mathrm{Pair}}; D_{\mathrm{BR}}, \ell)$, but the former requires the constraint $s_{\mathrm{Pair}} \in \mathcal{S}_{\mathrm{Decomp}}$. For the results of an unconstrained and constrained minimisation to coincide, at least one solution to the unconstrained minimisation must belong to the constraint set. Thus, the two Bayes-risks will coincide if and only if there is at least one minimiser of $\mathbb{L}(s_{\mathrm{Pair}}; D_{\mathrm{BR}}, \ell)$ that is in $\mathcal{S}_{\mathrm{Decomp}}$, i.e. $\ell \in \mathcal{L}_{\mathrm{Decomp}}$. ■

The condition in the right hand side above shall prove important in our subsequent analysis, so much so that we shall use $\mathcal{L}_{\mathrm{Decomp}}$ to denote the set of losses satisfying it:

$$\mathcal{L}_{\mathrm{Decomp}} \doteq \left\{\ell \mid (\forall D \in \Delta_{\mathcal{X}\times\{\pm 1\}})\, \mathcal{S}^*(D_{\mathrm{BR}}, \ell) \cap \mathcal{S}_{\mathrm{Decomp}} \neq \emptyset\right\}. \tag{20}$$

We shall revisit this condition when computing the Bayes-optimal scorers for the bipartite ranking risk. In particular, Proposition 44 characterises the set of proper composite losses for which $\ell \in \mathcal{L}_{\mathrm{Decomp}}$, which turns out to involve a condition on the link function for the loss.

5. The area under the ROC curve (AUC) and bipartite risk

The canonical performance measure for a scorer in bipartite ranking is the area under the ROC curve (AUC). In this section, we formally define the ROC curve and AUC, and show how the latter is related to the bipartite risk defined in Equation 16. We then establish several properties of the ROC curve and AUC, and contrast them to those of the classification risk for proper composite losses. We also describe a generalisation of the AUC based on our general bipartite risk. While some of these results are not new, presenting them together shall further delineate the distinctions between bipartite ranking and class-probability estimation.

Before describing the ROC curve, we must define the true and false positive rates for a scorer.

5.1 True and false positive rates

The goal of bipartite ranking is to produce a scorer $s\colon \mathcal{X} \to \mathbb{R}$. However, as many practical applications require a classifier $c\colon \mathcal{X} \to \{\pm 1\}$, it is of interest to convert a scorer into a classifier. The simplest approach to doing so is to *threshold* the scorer at some $t \in \mathbb{R}$, producing a classifier $c(x; t) = [\![s(x) \geq t]\!]$. One may assess the performance of the resulting classifier based on the *true* and *false positive rates*[8], which intuitively measure the accuracy (error) rates on each of the classes. These are defined below.

Definition 4 (True and false positive rates) *Given any $D = \langle P, Q, \pi\rangle \in \Delta_{\mathcal{X}\times\{\pm 1\}}$, and a scorer $s\colon \mathcal{X} \to \mathbb{R}$, define the true (false) positive (negative) rates of the scorer at a threshold $t \in \mathbb{R} \cup \{\pm\infty\}$ to be*

$$\mathrm{TPR}(t; D, s) \doteq \mathbb{P}_{\mathsf{X}\sim P}[s(\mathsf{X}) > t] + \frac{1}{2} \cdot \mathbb{P}_{\mathsf{X}\sim P}[s(\mathsf{X}) = t]$$

8. Depending on the literature, different terms may be used for these quantities. The true positive rate is sometimes referred to as the *recall* or *sensitivity*, and the true negative rate the *specificity*. The false positive rate is sometimes referred to as the *Type I error rate*, and the false negative rate the *Type II error rate*.

$$\mathrm{FPR}(t; D, s) \doteq \mathbb{P}_{\mathsf{X}' \sim Q}[s(\mathsf{X}') > t] + \frac{1}{2} \cdot \mathbb{P}_{\mathsf{X}' \sim Q}[s(\mathsf{X}') = t]$$
$$\mathrm{TNR}(t; D, s) \doteq 1 - \mathrm{FPR}(t; D, s) = \mathbb{P}_{\mathsf{X}' \sim Q}[s(\mathsf{X}') < t] + \frac{1}{2} \cdot \mathbb{P}_{\mathsf{X}' \sim Q}[s(\mathsf{X}') = t]$$
$$\mathrm{FNR}(t; D, s) \doteq 1 - \mathrm{TPR}(t; D, s) = \mathbb{P}_{\mathsf{X} \sim P}[s(\mathsf{X}) < t] + \frac{1}{2} \cdot \mathbb{P}_{\mathsf{X} \sim P}[s(\mathsf{X}) = t].$$

When the scorer s and distribution D are clear from context, we use e.g. $\mathrm{TPR}(t)$ *to denote* $\mathrm{TPR}(t; D, s)$.

In order to describe the properties of the true and false positive rates, it will be convenient to express them in terms of the distribution of the scorer. Denoting by P_S, Q_S the conditional distribution of scores on the positive and negative class when applied to instances drawn from D, we may write the true and false positive rates as:

$$\mathrm{TPR}(t; D, s) = \mathbb{P}_{\mathsf{S} \sim P_S}[\mathsf{S} > t] + \frac{1}{2} \cdot \mathbb{P}_{\mathsf{S} \sim P_S}[\mathsf{S} = t]$$
$$\mathrm{FPR}(t; D, s) = \mathbb{P}_{\mathsf{S}' \sim Q_S}[\mathsf{S}' > t] + \frac{1}{2} \cdot \mathbb{P}_{\mathsf{S}' \sim Q_S}[\mathsf{S}' = t].$$

The second term in each expression encodes that ties are broken uniformly at random between the positive and negative classes. Observe that the second term is zero unless there are isolated scores; i.e., the distribution of scores has a discrete random variable component. It is evident that both rates are non-increasing functions of t, and thus possess pseudo-inverses (Equation 1).

We are now in a position to describe the ROC curve and its basic properties.

5.2 The ROC curve and its properties

As discussed above, given a scorer $s : \mathcal{X} \to \mathbb{R}$, one may threshold it to produce a classifier. However, it is not *a priori* clear how to choose a suitable threshold. The ROC curve (Egan, 1975; Fawcett, 2006) is a graphical representation of a scorer that spells out the implications of every threshold choice. It is formed by tracing out the relationship between the true and false positive rates of a scorer across *all* possible thresholds.

Definition 5 (ROC curve) *Given any $D = \langle P, Q, \pi \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$, and a scorer $s : \mathcal{X} \to \mathbb{R}$, the ROC curve is defined by the parametric representation*[9]

$$\mathrm{ROC}(s; D) \doteq \{(\mathrm{FPR}(t; D, s), \mathrm{TPR}(t; D, s)) : t \in \mathbb{R} \cup \{\pm\infty\}\} \subseteq [0, 1]^2.$$

Equivalently, let $\rho(\alpha; D, s)$ be the power *of s at a false-positive rate of α, defined as*

$$(\forall \alpha \in [0, 1])\ \rho(\alpha; D, s) \doteq \mathrm{TPR}(\mathrm{FPR}^{-1}(\alpha)), \tag{21}$$

where we use the pseudo-inverse (Equation 1) of the false-positive rate. Then,

$$\mathrm{ROC}(s; D) = \{(\alpha, \rho(\alpha; D, s)) : \alpha \in \mathrm{Im}(\mathrm{FPR})\} \subseteq [0, 1]^2.$$

Put plainly, for all possible thresholds $t \in \mathbb{R}$, we create a classifier from s using the threshold, and assess the resulting accuracy on the positive and negative classes respectively. Every point on the ROC curve represents a pair of *realisable* true and false positive rates, i.e. rates that may be attained by a classifier based on an appropriate thresholding of the scorer. Thus, the ROC curve visually summarises the set of realisable error rates for various choices of classifier derived from the underlying scorer s.

In general, two points on the ROC curve represent different tradeoffs in terms of the true and false positive rates. Thus, there is no clear automated mechanism to pick the "best" threshold for generating a classifier without additional information as to one's underlying utility.

9. At a threshold $t = +\infty$, both the FPR and TPR are 0, while at $t = -\infty$, the FPR and TPR are both 1. Thus, varying t in this way traces a curve from left to right.

5.2.1 BASIC PROPERTIES OF THE ROC CURVE

We make some basic observations about the ROC curve for any scorer $s : \mathcal{X} \to \mathbb{R}$ and distribution D.

(i) we have $\{(0,0),(1,1)\} \subseteq \text{ROC}(s; D)$. This is because $\lim_{t\to\infty} \text{TPR}(t) = 0$, $\lim_{t\to-\infty} \text{TPR}(t) = 1$, and similarly for the FPR.

(ii) the curve does not have self-intersections. This is because FPR and TPR are monotone functions.

(iii) the curve is invariant to strictly monotone increasing transformations of the scorer s. This is because the FPR and TPR are invariant to strictly monotone increasing transformations (see also Proposition 15).

(iv) the curve is not necessarily surjective onto $[0,1]^2$, and may comprise only a finite number of points in $[0,1]^2$. This is a consequence of the fact that FPR and TPR are not necessarily *strictly* monotone.

(v) even if the curve is surjective, the set of points below the curve is not necessarily a convex set i.e. $\rho(\cdot)$ is not necessarily a concave function[10].

Figure 1 gives examples of various properties an ROC curve may possess, and in particular illustrates points (iv) and (v). We now discuss these last two points further.

5.2.2 INTERPOLATION OF THE ROC CURVE

In practice, one typically only has access to an empirical distribution $\hat{D}$ with finite support. The resulting empirical ROC curve will thus comprise a number of isolated points. In such situations, it is common to linearly interpolate between these points. To justify this interpolation, recall that each point on the ROC curve summarises the performance of a classifier derived from a specific threshold. Every interpolated point is similarly achievable by some *randomised* classifier derived from s (Scott et al., 1998, Theorem 1), (Provost and Fawcett, 2001, Theorem 7). To see why this is so, suppose $\text{ROC}(s; D) = \{(f_i, t_i)\}_{i=1}^n$, where $f_i \leq f_{i+1}, t_i \leq t_{i+1}$. For any $i \in [n-1]$, pick any $\alpha \in [0,1]$, and consider the interpolated point (f', t') defined by

$$(f', t') = \alpha \cdot (f_i, t_i) + (1-\alpha) \cdot (f_{i+1}, t_{i+1}).$$

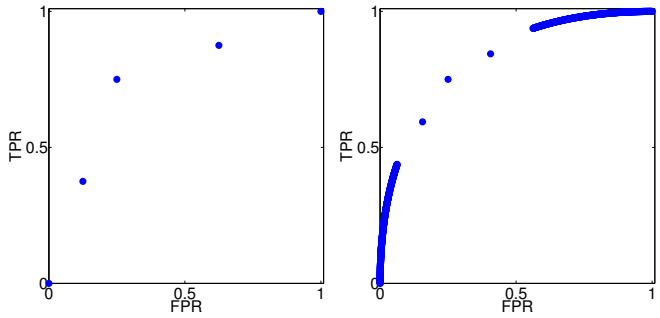
This corresponds exactly to the false positive and true positive rates of a randomised classifier derived from s using a threshold t_i, where ties are broken in favour of positives with probability α. Therefore, linear interpolation of the ROC curve summarises the performance of *all* possible classifiers with *randomised* tie-breaking that can be derived from the underlying scorer.

5.2.3 THE CONVEXIFIED ROC CURVE

To further build upon linear interpolation, one may construct a *convexified ROC curve* from the convex hull of $\text{ROC}(s; D)$[11] (Provost and Fawcett, 2001), which we denote by $\text{ROC}_{\text{cvx}}(s; D)$. As with the linearly interpolated ROC curve, every point on this curve is achievable with a suitable randomised classifier derived from the scorer (Provost and Fawcett, 2001, Theorem 7). It is not hard to justify the use of classifiers derived from $\text{ROC}_{\text{cvx}}(s; D)$, rather than those of $\text{ROC}(s; D)$: for every false positive rate, the classifiers derived from the former possess a true positive rate at least as good as that of a classifier derived from the latter. That is, the curve $\text{ROC}_{\text{cvx}}(s; D)$ *dominates* $\text{ROC}(s; D)$.

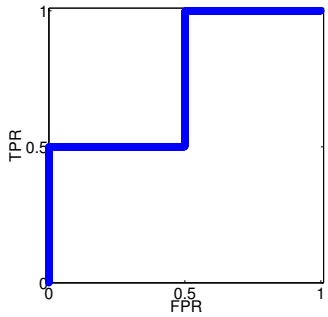
When $\text{ROC}(s; D)$ comprises a number of isolated points, $\text{ROC}_{\text{cvx}}(s; D)$ is trivially the linear interpolation of the ROC curve for a scorer whose realisable true and false positive rates are the hull points. Such a scorer may be derived from s by introducing suitable ties. This process can be shown to be equivalent to performing an *isotonic regression* (Ayer et al., 1955) on the original scorer s (Fawcett and Niculescu-Mizil, 2007).

10. Recall a function is convex iff its epigraph is a convex set, and is concave iff its negation is convex.
11. Strictly, if $\text{co}(\cdot)$ denotes the convex hull of a set, one considers $\text{ROC}_{\text{cvx}}(s; D) = \{(0,0),(1,1)\} \cup (\text{co}(\text{ROC}(s; D)) \cap \{(\alpha, \rho) \mid \rho > \alpha\})$, so that only the portion of the convex hull strictly above the diagonal line is retained. When the original curve is entirely below the diagonal line, one just uses the diagonal line itself.

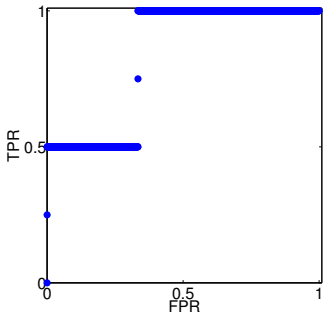


(a) Isolated points.

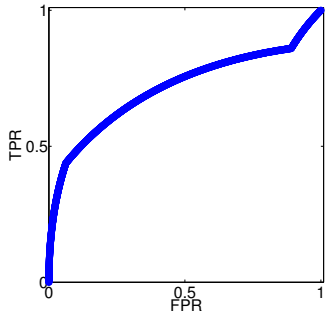
(b) Disconnected.



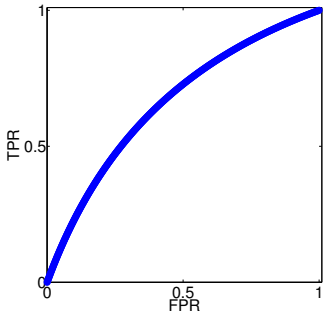
(c) Vertical steps.



(d) Discontinuous.



(e) Continuous but non-concave.



(f) Concave and differentiable.

Figure 1: Illustration of various properties an ROC curve may possess. In each example, we use $\mathcal{X} = [0, 1]$, and in most cases fix $\eta(x) = x$. For (a), we use the scorer $s = \text{sign}(2 \cdot \eta - 1)$, so that the score distributions P_S, Q_S are discrete. For (b), we use the scorer $s(x) = \eta(x) \vee 0.75$ when $\eta(x) < 0.5$ and $s(x) = \eta(x) \wedge 0.25$ else, so that the score distributions have isolated areas. For (c), we round η to $\{0, 1\}$ and use the scorer s predicting uniformly 0.5 or 1 when $\eta(x) > 0.5$ and $\frac{3}{8} \cdot \eta(x)$ else, so that the positive score distribution has isolated elements. For (d), we round η to $\{0, 1\}$ and use the scorer s predicting uniformly in $[0.25, 0.5] \cup [0.9, 1]$ when $\eta(x) > 0.5$ and $[0, 0.25] \cup [0.5, 0.75]$ else, so that the score distributions have interleaved support. For (e), we use the scorer $s(x) = \eta(x) - 0.5$ when $\eta(x) < 0.5$ and $s(x) = 4 \cdot \eta(x) - 3$ else, so that the score distributions are overlapping. For (f), we use the scorer $s = \eta$, so that the result is the maximal ROC curve.

5.2.4 EXTREMAL ROC CURVES

Given a distribution D, different scorers will induce different ROC curves. It is natural to ask whether there is a "best" possible curve for a given D, that is, a curve that *dominates* every other one. It turns out that there is such a curve, corresponding to any scorer which is a strictly monotone increasing transformation of the class-probability function η. This fact is a consequence of the classical Neyman-Pearson lemma, which plays an important role in hypothesis testing. Appendix D provides some background on this lemma.

Lemma 6 *Given any distribution $D = \langle M, \eta \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$ and scorer $s : \mathcal{X} \to \mathbb{R}$,*

$$(\forall \alpha \in [0, 1])\, \rho(\alpha; D, s) \leq \rho(\alpha; D, \phi \circ \eta)$$

where ϕ is any strictly monotone increasing function.

Proof By the Neyman-Pearson lemma, the uniformly most powerful test (i.e. the test with maximal $\rho(\cdot)$ value) at any α is given by the likelihood ratio. For the distribution D, this is nothing but a strictly monotone transformation of η. As the ROC curve is invariant to such transformations, the result follows. ■

We may equally consider the "worst" possible curve for a given D, that is, a curve that is *dominated* by every other one. Since the curve for a scorer s has as mirror image the curve for the scorer $-s$, evidently, such a curve will correspond to any strictly monotone *decreasing* function of η, such as e.g. $s = -\eta$.

Finally, it is easy to check that a non-informative scorer s that uniformly predicts a constant (e.g. $s \equiv 0$) will induce a ROC curve that is the diagonal, viz. $\rho(\alpha) = \alpha$. Thus, intuitively, a scorer must induce a ROC curve significantly away from the diagonal in order to be useful.

5.2.5 DIFFERENTIABILITY AND CONCAVITY OF THE ROC CURVE

We now consider properties of the derivative of the ROC curve, when it exists. When the curve comprises a number of isolated points, then the derivatives of the interpolated version of the curve can be easily computed. More generally, from the definition of the power function (Equation 21), it is apparent that the differentiability of the ROC curve relies on that of the true- and false-positive rates. The following makes this precise.

Proposition 7 ((Clémençon and Vayatis, 2009, Proposition 24)) *Given a distribution $D = \langle M, \eta \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$ with M absolutely continuous and scorer $s : \mathcal{X} \to \mathbb{R}$ with corresponding induced class-conditional distributions P_S, Q_S, the curve* $\mathrm{ROC}(s; D)$ *is differentiable if and only if P_S, Q_S are absolutely continuous.*

While in practice one's operating P_S, Q_S may be discrete, it is common to construct continuous approximations to them e.g. the binormal ROC model (Krzanowski and Hand, 2009, Section 2.5). Observe that when the false and true positive rates are differentiable, we have

$$\begin{aligned} (\forall t \in \mathbb{R})\, \mathrm{TPR}'(t) &= -p_S(t) \\ \mathrm{FPR}'(t) &= -q_S(t), \end{aligned} \tag{22}$$

where p_S, q_S are the densities of the class-conditional score distributions. When the ROC curve is differentiable, the derivative turns out to have a well-known form involving the score-to-probability transform of the scorer, as stated below. (Appendix C gives a proof for completeness.)

Proposition 8 ((Krzanowski and Hand, 2009, pg. 22), (Clémençon and Vayatis, 2009b, Lemma 1)) *Given a distribution $D = \langle P, Q, \pi \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$ and scorer $s : \mathcal{X} \to \mathbb{R}$ such that* $\mathrm{ROC}(s; D)$ *is differentiable and there exist induced class-conditional densities p_S, q_S, the slope of the ROC curve at a false positive rate $\alpha \in (0, 1)$ is*

$$\begin{aligned} \rho'(\alpha) &= \frac{p_S(\mathrm{FPR}^{-1}(\alpha))}{q_S(\mathrm{FPR}^{-1}(\alpha))} \\ &= \frac{1 - \pi}{\pi} \cdot \frac{\mathrm{Prb}(\mathrm{FPR}^{-1}(\alpha))}{1 - \mathrm{Prb}(\mathrm{FPR}^{-1}(\alpha))}. \end{aligned} \tag{23}$$

Proposition 8 establishes that the score-to-probability transform $\mathrm{Prb}(\cdot\,; D, s)$, and thus the calibration transform $\mathrm{Cal}(\cdot\,; D, s)$, may be computed based on the derivatives of the ROC curve. The fact that the ROC curve may be used to obtain a class-probability estimator is well-known (Fawcett and Niculescu-Mizil, 2007; Flach, 2010). Indeed, the relationship implies a characterisation of the monotonicity of the calibration transform when the ROC curve is differentiable.

Corollary 9 *Given a distribution $D \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ and scorer $s\colon \mathcal{X} \to \mathbb{R}$ with differentiable ROC curve and invertible rates,* $\mathrm{ROC}(s; D)$ *is (strictly) concave if and only if the function* $\mathrm{Prb}(\cdot\,; D, s)$ *is (strictly) monotone.*

Proof By Proposition 8, the derivative of the curve at any point comprises a strictly monotone transform of Prb composed with the invertible and hence strictly monotone function FPR^{-1}. As (strict) concavity of the curve is equivalent to (strict) monotonicity of its derivative, the result follows. ■

Non-concave regions of the ROC curve thus correspond to regions where the function Prb is non-monotone, and hence non-invertible. As we have seen, one may compute the convex hull of the ROC curve to remove such regions, which corresponds to introducing ties in the scorer.

When the scorer s is calibrated (Equation 5), Equation 23 implies that the slope of $\mathrm{ROC}(s; D)$ at a false positive rate $\alpha \in [0, 1]$ simplifies to

$$\rho'(\alpha) = \frac{1-\pi}{\pi} \cdot \frac{\mathrm{FPR}^{-1}(\alpha)}{1 - \mathrm{FPR}^{-1}(\alpha)}. \tag{24}$$

Further, Proposition 8 implies the following useful fact.

Corollary 10 *Given a distribution $D = \langle P, Q, \pi\rangle \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ and calibrated scorer $s\colon \mathcal{X} \to \mathbb{R}$ with differentiable ROC curve,* $\mathrm{ROC}(s; D)$ *is strictly concave.*

Proof As s is calibrated, by definition $\mathrm{Prb}(\cdot\,; D, s)$ is the identity mapping, and thus strictly monotone. By Corollary 9 the result follows. ■

This implies that for the calibrated scorer $s = \eta$, the ROC curve is concave, as noted before in e.g. (Clémençon and Vayatis, 2009, Proposition 8). We note that Corollary 10 relies crucially on the ROC curve being differentiable: the trivially calibrated scorer $s \equiv \pi$ would have a non-differentiable ROC curve comprising isolated points, whose interpolation would not be strictly concave.

5.2.6 THE ROC CURVE AND COST-SENSITIVE THRESHOLD SELECTION

The true positive and negative rates measure the accuracy of a classifier on the positive and negative classes respectively. Observe that the standard 0-1 classification risk (Equation 15) can be expressed in terms of the false positive and negative rates using a threshold of 0. More generally, given a cost ratio $c \in [0, 1]$, the *cost-sensitive risk* of a scorer $s\colon \mathcal{X} \to \mathbb{R}$ when using a threshold t may be written

$$\mathbb{L}(s; D, \ell_{\mathrm{CS}(c,t)}) = \pi \cdot (1-c) \cdot \mathrm{FNR}(t; D, s) + (1-\pi) \cdot c \cdot \mathrm{FPR}(t; D, s), \tag{25}$$

where $\ell_{\mathrm{CS}(c,t)}$ denotes the cost-sensitive loss with cost ratio c and threshold choice t:

$$\begin{aligned} \ell_{\mathrm{CS}(c,t)}(+1, z) &\doteq (1-c) \cdot \ell_{01}(+1, z-t) \\ \ell_{\mathrm{CS}(c,t)}(-1, z) &\doteq c \cdot \ell_{01}(-1, z-t). \end{aligned} \tag{26}$$

Clearly, $\ell_{\mathrm{CS}(c,c)} = \ell_{\mathrm{CS}(c)}$, and for $t = 0$ and $c = \frac{1}{2}$ we recover the (scaled) 0-1 loss. The *optimal threshold function* $t^*\colon [0, 1] \mapsto 2^{\mathbb{R}}$ maps costs to the set of thresholds yielding minimal cost-sensitive risk:

$$(\forall c \in [0, 1])\, t^*(c; D, s) = \operatorname*{Argmin}_{t\in\mathbb{R}} \mathbb{L}(s; D, \ell_{\mathrm{CS}(c,t)}).$$

Determining the optimal threshold function for a scorer is intimately related to the gradient of the ROC curve. (Appendix C provides a proof for completeness.)

Proposition 11 ((Krzanowski and Hand, 2009, pg. 24)) *Given any $D = \langle P, Q, \pi \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$ and scorer $s : \mathcal{X} \to \mathbb{R}$ with differentiable ROC curve, for any cost ratio $c \in (0, 1)$, any optimal threshold $t_0 \in t^*(c; D, s)$ for the cost-sensitive risk $\mathbb{L}(s; D, \ell_{\mathrm{CS}(c,t)})$ satisfies*

$$\rho'(\mathrm{FPR}(t_0)) = \frac{c}{1-c} \cdot \frac{1-\pi}{\pi},$$

or equivalently,

$$\mathrm{Prb}(t_0) = c.$$

Further, when the score-to-probability mapping $\mathrm{Prb}(\cdot\,; D, s)$ is invertible,

$$t^*(c) = \{\mathrm{Prb}^{-1}(c)\}.$$

Intuitively, there are multiple optimal thresholds for a given cost c when $\mathrm{Prb}(\cdot\,; D, s)$ is *not* invertible, i.e. when the ROC curve is not strictly concave. The derivative may also have an image that is a strict subset of $\mathbb{R}_+$. In this case, the risk in Equation 25 is monotone as a function of t, and so the optimal threshold is one of $\pm\infty$.

The relationship between the slope of the ROC and the optimal threshold is useful in two ways. First, given a particular c, to find the optimal threshold, we draw a line of slope $\frac{c}{1-c} \cdot \frac{1-\pi}{\pi}$. The point at which it touches the ROC curve corresponds to the optimal threshold. Second, at a given false positive rate α achieved by some threshold, the derivative of the ROC curve gives us the cost for which the given threshold is optimal.

For the case of calibrated scorers, we have a simpler characterisation of the optimal threshold for a cost-sensitive loss: it is simply the corresponding cost itself.

Corollary 12 *Given any $D = \langle P, Q, \pi \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$ and calibrated scorer $s : \mathcal{X} \to \mathbb{R}$ with differentiable ROC curve, for any cost $c \in (0, 1)$, the optimal threshold function for the cost-sensitive risk is $t^*(c) = \{c\}$.*

Proof As s is calibrated, $\mathrm{Prb}(\cdot\,; D, s)$ is the identity mapping, and thus invertible. Thus, applying Proposition 11, we know there is a unique optimal threshold $t_0(c)$ for each cost ratio c. Applying Equation 24, the optimal threshold satisfies

$$\frac{t_0(c)}{1 - t_0(c)} = \frac{c}{1-c},$$

i.e. the optimal threshold is $t_0(c) = c$. ■

Corollary 12 is again evident for the trivially calibrated scorer $s = \eta$.

5.2.7 ROC DOMINATION AND CLASSIFICATION RISK

Suppose one scorer dominates another in ROC space. What can one say about the risks of the two scorers with respect to a proper loss? The following establishes that, for calibrated scorers, dominance in ROC space implies dominance with respect to *any* proper composite risk. This supports the use of the ROC curve as a means of assessing the performance of a scorer. While simple to prove, the result is to our knowledge novel.

Proposition 13 *Pick any distribution $D \in \Delta_{\mathcal{X} \times \{\pm 1\}}$, and let $s_1, s_2 : \mathcal{X} \to \mathbb{R}$ be any scorers that are calibrated with respect to D, with differentiable ROC curves. If the ROC curve of s_2 dominates that of s_1,*

$$(\forall \alpha \in [0, 1])\ \rho(\alpha; D, s_1) \leq \rho(\alpha; D, s_2).$$

then any *proper loss λ, s_2 must have risk no larger than s_1:*

$$\mathbb{L}(s_2; D, \lambda) \leq \mathbb{L}(s_1; D, \lambda).$$

Proof Our basic idea will be to show that s_2 has a lower cost-sensitive risk than s_1 for any conceivable cost-ratio, which by Shuford's representation (Equation 10) will yield the desired result. For brevity, in the following we use e.g. $\mathrm{TPR}_s(c)$ as a shorthand for $\mathrm{TPR}(c; D, s)$.

Pick any $c \in (0, 1)$. Consider the cost-sensitive risk of a scorer s using a threshold t (Equation 25). Since s_1, s_2 are calibrated, by Corollary 12, they both have optimal threshold $t^*(c; D, s) = \{c\}$. With this threshold, scorer s_1 achieves a false-positive rate of $\alpha_1 \doteq \mathrm{FPR}_{s_1}(c)$, and true-positive rate $\mathrm{TPR}_{s_1}(c)$. By the ROC domination assumption, we have

$$\mathrm{TPR}_{s_1}(c) = \rho(\alpha_1; D, s_1) \leq \rho(\alpha_1; D, s_2) = \mathrm{TPR}_{s_2}((\mathrm{FPR}_{s_2})^{-1}(\alpha_1)).$$

Thus, the corresponding false negative rate of s_2 must be smaller than that of s_1, i.e.

$$\mathrm{FNR}_{s_2}((\mathrm{FPR}_{s_2})^{-1}(\alpha_1)) \leq \mathrm{FNR}_{s_1}(c).$$

This implies that using a threshold of $t_2 = (\mathrm{FPR}_{s_2})^{-1}(\alpha_1)$, s_2 achieves a lower cost-sensitive risk than s_1:

$$\begin{aligned}
\mathbb{L}(s_2; \ell_{\mathrm{CS}(c,t_2)}) &= \pi \cdot (1-c) \cdot \mathrm{FNR}_{s_2}(t_2) + (1-\pi) \cdot c \cdot \mathrm{FPR}_{s_2}(t_2) \\
&\leq \pi \cdot (1-c) \cdot \mathrm{FNR}_{s_1}(c) + (1-\pi) \cdot c \cdot \alpha_1 \\
&= \pi \cdot (1-c) \cdot \mathrm{FNR}_{s_1}(c) + (1-\pi) \cdot c \cdot \mathrm{FPR}_{s_1}((\mathrm{FPR}_{s_1})^{-1}(\alpha_1)) \\
&= \mathbb{L}(s_1; c, \ell_{\mathrm{CS}(c)}).
\end{aligned}$$

But since c is the optimal threshold for s_2, we further have

$$\begin{aligned}
\mathbb{L}(s_2; \ell_{\mathrm{CS}(c)}) &= \mathbb{L}(s_2; \ell_{\mathrm{CS}(c,c)}) \\
&\leq \mathbb{L}(s_2; \ell_{\mathrm{CS}(c,t_2)}) \\
&\leq \mathbb{L}(s_1; \ell_{\mathrm{CS}(c,c)}) \\
&= \mathbb{L}(s_1; \ell_{\mathrm{CS}(c)}).
\end{aligned}$$

Since s_2 has a lower cost-sensitive risk than s_1 for any cost ratio c, by Shuford's representation (Equation 10), it must also have lower risk with respect to any proper loss. ■

An immediate consequence of the above is that for scorers with strictly monotone calibration transforms $\mathrm{Cal}(\cdot\,; D, s)$ (in turn relying on strict concavity of the ROC curve, by Corollary 10), ROC dominance implies dominance with respect to any proper loss. Note also that for the optimal ROC curve, given by $s^* = \eta$, the above is trivially true as any proper loss will have its risk minimised by exactly η.

Proposition 13 relies on ROC dominance. When the ROC curves for two scorers cross, it is not hard to construct examples of losses where one scorer is superior to the other. This indicates that in such situations, caution must be used before declaring one scorer to be superior to another, as is well known (Hand, 2009).

5.3 The area under the ROC curve (AUC)

The ROC curve is a graphical display of the performance of a scorer. It is often desirable to additionally have a single numeric summary of performance. One such popular summary statistic is the *area under the ROC curve*, or *AUC*.

Definition 14 (Area under the ROC curve (AUC)) *Given any $D \in \Delta_{\mathcal{X} \times \{\pm 1\}}$ and scorer $s : \mathcal{X} \to \mathbb{R}$, the* area under the ROC curve *or* AUC *of s is the area under the curve* $\mathrm{ROC}(s; D)$*,*

$$\mathrm{AUC}(s; D) \doteq \int_0^1 \rho(\alpha; D, s)\, d\alpha, \tag{27}$$

where $\rho(\alpha; D, s)$ is the power *of s at α (Equation 21).*

A subtlety in the above definition is that we only defined $\mathrm{ROC}(s; D)$ in terms of the power when $\alpha \in \mathrm{Im}(\mathrm{FPR})$. However, the power itself is defined for every $\alpha \in [0, 1]$, due to the use of the pseudo-inverse. Thus, the integral is well-defined even when $\mathrm{ROC}(s; D)$ comprises isolated points.

A further subtlety is that the curve traced out by $\{(\alpha, \mathrm{TPR}(\mathrm{FPR}^{-1}(\alpha))) : \alpha \in [0, 1]\}$ is *not* always equivalent to that generated by linear interpolation of $\mathrm{ROC}(s; D)$. Nonetheless, the area under the two curves will in general be the same. To see this, suppose P_S, Q_S have an isolated component at some $t \in \mathbb{R}$, with $\mathrm{FPR}(t^-) = \alpha_1, \mathrm{FPR}(t^+) = \alpha_2$ and $\mathrm{TPR}(t^-) = \beta_1, \mathrm{TPR}(t^+) = \beta_2$, and further $\mathrm{TPR}(t) = \frac{1}{2}(\beta_1 + \beta_2)$ due to the breaking of ties uniformly at random. We will then have two consecutive disconnected points in $\mathrm{ROC}(s; D)$, (α_1, β_1) and (α_2, β_2). With linear interpolation, this region of the ROC curve has area $\frac{1}{2} \cdot (\alpha_2 - \alpha_1) \cdot (\beta_1 + \beta_2)$. For any $\alpha \in (\alpha_1, \alpha_2)$, we have $(\mathrm{FPR})^{-1}(\alpha) = t$, and so $\mathrm{TPR}((\mathrm{FPR})^{-1}(\alpha)) = \mathrm{TPR}(t) = \frac{1}{2}(\beta_1 + \beta_2)$. Thus with the pseudo-inverse, the corresponding area of this region is that of the corresponding rectangle with height $\frac{1}{2}(\beta_1 + \beta_2)$ and width $(\alpha_2 - \alpha_1)$, which is also exactly $\frac{1}{2} \cdot (\alpha_2 - \alpha_1) \cdot (\beta_1 + \beta_2)$.

5.3.1 BASIC PROPERTIES OF THE AUC

Some basic properties of the AUC are immediate from the above definition. First, the AUC is in $[0, 1]$; as we shall subsequently discuss, higher AUC values indicate a "better" scorer.

Second, the AUC is independent of the base rate π, and only depends on the class-conditional distributions P, Q for a distribution $D \in \Delta_{\mathcal{X} \times \{\pm 1\}}$. This means that a scorer s will have the same AUC with respect to all distributions in the family $\{D = \langle P, Q, \pi \rangle\}_{\pi \in [0,1]}$.

Third, the AUC is invariant to strictly monotone increasing transforms of the scorer, as shown below.

Proposition 15 ((Clémençon and Vayatis, 2009, Proposition 24)) *Given any $D \in \Delta_{\mathcal{X} \times \{\pm 1\}}$ and scorer $s : \mathcal{X} \to \mathbb{R}$, for any strictly monotone increasing $\phi : \mathbb{R} \to \mathbb{R}$,*

$$\mathrm{AUC}(s; D) = \mathrm{AUC}(\phi \circ s; D).$$

Proof This is because the 0-1 loss is invariant to monotone increasing transformations of the scorer:

$$\begin{aligned}(\forall t \in \mathbb{R})\, \mathrm{FPR}(t; D, \phi \circ s) &= \mathop{\mathbb{E}}_{\mathsf{X} \sim Q}[\ell_{01}(-1, \phi(s(\mathsf{X})) - t)] \\ &= \mathop{\mathbb{E}}_{\mathsf{X} \sim Q}[\ell_{01}(-1, s(\mathsf{X}) - t)] \\ &= \mathrm{FPR}(t; D, s),\end{aligned}$$

and similarly for TPR. It follows that the power function ρ is unaffected by ϕ, and thus so is the ROC curve and the area under it. ■

Fourth, the AUC is optimised by any strictly monotone transformation of the underlying η.

Corollary 16 *For any $D = \langle M, \eta \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$ and any strictly monotone increasing $\phi : [0, 1] \to \mathbb{R}$,*

$$\sup_{s :\, \mathcal{X} \to \mathbb{R}} \mathrm{AUC}(s; D) = \mathrm{AUC}(\phi \circ \eta; D).$$

Proof By Lemma 6, the optimal ROC curve is achieved by any $\phi \circ \eta$. Such a scorer must thus have maximal AUC. ■

We now show how the AUC can be viewed in terms of loss functions, which makes apparent its connection to the bipartite ranking risk.

5.3.2 A LOSS REPRESENTATION OF THE AUC

Although not immediately obvious from the above definition, the AUC of a scorer with respect to a distribution $D \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ is the probability a randomly drawn positive has a higher score than a randomly drawn negative, with ties broken uniformly at random. This observation goes back to at least Hanley and McNeil (1982, Section III), and has been noted in machine learning community in e.g. Cortes and Mohri (2003, Lemma 1), Clémençon et al. (2008).

Proposition 17 ((Cortes and Mohri, 2003, Lemma 1), (Clémençon et al., 2008)) [12] *Given any* $D = \langle P, Q, \pi\rangle \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ *and scorer* $s : \mathcal{X} \to \mathbb{R}$,

$$\mathrm{AUC}(s; D) = \mathbb{P}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}[s(\mathsf{X}) > s(\mathsf{X}')] + \frac{1}{2}\cdot \mathbb{P}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}[s(\mathsf{X}) = s(\mathsf{X}')]. \tag{28}$$

Equation 28 is often taken as the starting definition of the AUC, due to its convenience to manipulate. Indeed, on a sample $\hat{D} = \{(x_i, +1)\}_{i=1}^n \cup \{(x_j, -1)\}_{j=1}^m$, the empirical AUC is

$$\mathrm{AUC}(s; \hat{D}) = \frac{1}{n\cdot m}\cdot \sum_{i=1}^{n}\sum_{j=1}^{m}\left[\llbracket s(x_i) > s(x_j)\rrbracket + \frac{1}{2}\cdot\llbracket s(x_i) = s(x_j)\rrbracket\right],$$

which is equivalent to the Mann-Whitney statistic (Mann and Whitney, 1947). The resulting empirical estimate can be computed in $O(N \log N)$ time for $N = n + m$ with a single sort operation, rather than attempting numerical integration of the empirical ROC curve (Hand and Till, 2001).

Observe that Proposition 17 may be expressed in terms of a risk involving 0-1 loss as follows.

Corollary 18 *Given any* $D = \langle P, Q, \pi\rangle \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ *and scorer* $s : \mathcal{X} \to \mathbb{R}$,

$$\begin{aligned}\mathrm{AUC}(s; D) &= 1 - \mathop{\mathbb{E}}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[\llbracket s(\mathsf{X}) - s(\mathsf{X}') < 0\rrbracket + \frac{1}{2}\llbracket s(\mathsf{X}) = s(\mathsf{X}')\rrbracket\right]\\ &= 1 - \mathop{\mathbb{E}}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[\ell_{01}(1, s(\mathsf{X}) - s(\mathsf{X}'))\right].\end{aligned} \tag{29}$$

Building on this representation, we now describe a generalisation of the AUC, which will be a useful basis for further analysis.

5.4 Generalisation: the ℓ-AUC

We define the following generalisation of the AUC, which uses any loss function $\ell : \{\pm 1\}\times\mathbb{R} \to \mathbb{R}_+$.

Definition 19 (ℓ-AUC) *Given any* $D = \langle P, Q, \pi\rangle \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ *and a loss* $\ell : \{\pm 1\}\times\mathbb{R} \to \mathbb{R}_+$, *define the* ℓ-AUC *of a scorer* s *with respect to* $D \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ *as*

$$\mathrm{AUC}(s; D, \ell) \doteq 1 - \mathop{\mathbb{E}}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[\ell_{\mathrm{symm}}(s(\mathsf{X}) - s(\mathsf{X}'))\right],$$

recalling that ℓ_{symm} *is the symmetrised version of* ℓ *(Equation 8). When* ℓ *is symmetric, this simplifies to*

$$\mathrm{AUC}(s; D, \ell) = 1 - \mathop{\mathbb{E}}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[\ell_1(s(\mathsf{X}) - s(\mathsf{X}'))\right].$$

Table 6 provides some examples of the ℓ-AUC. Clearly, $\mathrm{AUC}(s; D, \ell_{01})$ is the standard AUC as defined earlier (see e.g. Equation 28). When we do not explicitly refer to the loss, it is understood that we are referring to the standard AUC.

12. Compared to the cited works, we have an extra term that accounts for ties.

ℓ	$\mathrm{AUC}(s; D, \ell)$
ℓ_{01}	$1 - \mathbb{E}_{\mathsf{X} \sim P, \mathsf{X}' \sim Q} \left[[\![s(\mathsf{X}) < s(\mathsf{X}')]\!] + \frac{1}{2} \cdot [\![s(\mathsf{X}) = s(\mathsf{X}')]\!] \right]$
ℓ_{sq}	$1 - \mathbb{E}_{\mathsf{X} \sim P, \mathsf{X}' \sim Q} \left[(1 - s(\mathsf{X}) + s(\mathsf{X}'))^2 \right]$
$\ell_{\log}$	$1 - \mathbb{E}_{\mathsf{X} \sim P, \mathsf{X}' \sim Q} \left[\log \left(1 + e^{s(\mathsf{X}') - s(\mathsf{X})} \right) \right]$
$\ell_{\exp}$	$1 - \mathbb{E}_{\mathsf{X} \sim P, \mathsf{X}' \sim Q} \left[e^{s(\mathsf{X}') - s(\mathsf{X})} \right]$

Table 6: Examples of ℓ-AUC for various losses, given some $D = \langle P, Q, \pi \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$.

When using the cost-sensitive loss $\ell_{\mathrm{CS}(c,t)}$ of Equation 26 with a threshold of $t = 0$ and any $c \in (0, 1)$, we also recover the standard AUC. This is because the symmetrised version of such a loss is

$$\begin{aligned} \ell_{\mathrm{CS}(c,0)}(1, v) + \ell_{\mathrm{CS}(c,0)}(-1, -v) &= (1 - c) \cdot [\![v < 0]\!] + c \cdot [\![-v > 0]\!] + \frac{1}{2} \cdot [\![v = 0]\!] \\ &= (1 - c) \cdot [\![v < 0]\!] + c \cdot [\![v < 0]\!] + \frac{1}{2} \cdot [\![v = 0]\!] \\ &= [\![v < 0]\!] + \frac{1}{2} \cdot [\![v = 0]\!] \\ &= \ell_{01}(1, v) \end{aligned}$$

This is intuitively because of the symmetry inherent in the bipartite ranking problem: scoring a positive below a negative is equivalent to scoring a negative above a positive. Therefore, one cannot expect to have different costs associated with the two errors.

We have assumed the prediction space for the loss ℓ above to $\mathbb{R}$ because we require the prediction space to be closed under negation, and for an arbitrary scorer that requires we can compute the loss for any real valued prediction. This rules out using a proper loss (or indeed any probability estimation loss), which is only defined on the prediction space $[0, 1]$. However, we may use a proper composite loss, which operates on a real-valued prediction space but converts this to $[0, 1]$ via a link function.

5.4.1 THE ℓ-AUC AS AN AREA

One property of the AUC that is inherited is that the ℓ-AUC may be interpreted as the area under a curve parameterised by the false positive rate, and a smoothed version of the true positive and false positive rates, defined below.

Definition 20 *Given any $D = \langle P, Q, \pi \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$, loss ℓ, and a scorer $s : \mathcal{X} \to \mathbb{R}$, define the ℓ-true (false) positive (negative) rates at a threshold $t \in \mathbb{R}$ to be*

$$\begin{aligned} \mathrm{FNR}(t; D, s, \ell) &= \mathop{\mathbb{E}}_{\mathsf{X} \sim P} [\ell_1(s(\mathsf{X}) - t)] \\ \mathrm{FPR}(t; D, s, \ell) &= \mathop{\mathbb{E}}_{\mathsf{X} \sim Q} [\ell_{-1}(s(\mathsf{X}) - t)] \\ \mathrm{TPR}(t; D, s, \ell) &= 1 - \mathrm{FNR}(t; D, s, \ell) \\ \mathrm{TNR}(t; D, s, \ell) &= 1 - \mathrm{FPR}(t; D, s, \ell). \end{aligned}$$

When the scorer s and distribution D are clear from context, we shall use e.g. $\mathrm{TPR}_\ell(t)$ to denote $\mathrm{TPR}(t; D, s, \ell)$.

We then have the following analogue to Equation 27 for a general loss.

Proposition 21 *Given any $D \in \Delta_{\mathcal{X}\times\{\pm 1\}}$, loss ℓ, and scorer $s : \mathcal{X} \to \mathbb{R}$ with differentiable ROC curve and invertible false- and true-positive rates*[13]*,*

$$\mathrm{AUC}(s; D, \ell) = \int_0^1 \rho(\alpha; D, s, \ell)\, d\alpha, \tag{30}$$

where $\rho(\alpha; D, s, \ell)$ is the ℓ-power *of s at α, defined for $\alpha \in [0, 1]$ as*

$$\rho(\alpha; D, s, \ell) = \frac{1}{2} \cdot \left(\mathrm{TPR}_\ell(\mathrm{FPR}^{-1}(\alpha)) + \mathrm{TNR}_\ell(\mathrm{TPR}^{-1}(\alpha))\right). \tag{31}$$

When ℓ is symmetric, this simplifies to

$$\mathrm{AUC}(s; D, \ell) = \int_0^1 \mathrm{TPR}_\ell(\mathrm{FPR}^{-1}(\alpha))\, d\alpha = \int_0^1 \mathrm{TNR}_\ell(\mathrm{TPR}^{-1}(\alpha))\, d\alpha.$$

Proof The proof is a generalisation of e.g. (Cortes and Mohri, 2003, Lemma 1), (Clémençon et al., 2008, Proposition B.2) to cover an arbitrary loss $\ell : \{\pm 1\} \times \mathbb{R} \to \mathbb{R}_+$. Recall that for any $t \in \mathbb{R}$,

$$\begin{aligned}
\mathrm{TPR}_\ell(t) &= 1 - \mathop{\mathbb{E}}_{\mathsf{X}\sim P}[\ell_1(s(\mathsf{X}) - t)] \\
\mathrm{TNR}_\ell(t) &= 1 - \mathop{\mathbb{E}}_{\mathsf{X}\sim Q}[\ell_{-1}(s(\mathsf{X}) - t)].
\end{aligned}$$

Starting from the definition, and by swapping the expectation and integral (which is justified by Tonelli's theorem (Folland, 1999, pg. 67), since ℓ_1 is nonnegative and measurable):

$$\begin{aligned}
\int_0^1 \mathrm{TPR}_\ell(\mathrm{FPR}^{-1}(\alpha))\, d\alpha &= 1 - \int_0^1 \mathop{\mathbb{E}}_{\mathsf{X}\sim P}[\ell_1(s(\mathsf{X}) - \mathrm{FPR}^{-1}(\alpha))]\, d\alpha \\
&= 1 - \mathop{\mathbb{E}}_{\mathsf{X}\sim P}\left[\int_0^1 \ell_1(s(\mathsf{X}) - \mathrm{FPR}^{-1}(\alpha))\, d\alpha\right] \\
&= 1 - \mathop{\mathbb{E}}_{\mathsf{X}\sim P}\left[-\int_{-\infty}^{\infty} \ell_1(s(\mathsf{X}) - t) \cdot \mathrm{FPR}'(t)\, dt\right] \text{ with } \alpha = \mathrm{FPR}(t) \\
&= 1 - \mathop{\mathbb{E}}_{\mathsf{X}\sim P}\left[\int_{-\infty}^{\infty} \ell_1(s(\mathsf{X}) - t) \cdot q_S(t)\, dt\right] \\
&= 1 - \mathop{\mathbb{E}}_{\mathsf{X}\sim P}\left[\int_{-\infty}^{\infty} \ell_1(s(\mathsf{X}) - t) \cdot \mathop{\mathbb{E}}_{\mathsf{X}'\sim Q}\left[\delta_{s(\mathsf{X}')}(t)\right]\, dt\right] \\
&= 1 - \mathop{\mathbb{E}}_{\mathsf{X}\sim P, \mathsf{X}'\sim Q}\left[\ell_1(s(\mathsf{X}) - s(\mathsf{X}'))\right],
\end{aligned}$$

where the last line follows from the definition of Dirac delta. Similarly,

$$\begin{aligned}
\int_0^1 \mathrm{TNR}_\ell(\mathrm{TPR}^{-1}(\alpha))\, d\alpha &= 1 - \int_0^1 \mathop{\mathbb{E}}_{\mathsf{X}'\sim Q}[\ell_{-1}(s(\mathsf{X}') - \mathrm{TPR}^{-1}(\alpha))]\, d\alpha \\
&= 1 - \mathop{\mathbb{E}}_{\mathsf{X}'\sim Q}\left[\int_0^1 \ell_{-1}(s(\mathsf{X}') - \mathrm{TPR}^{-1}(\alpha))\, d\alpha\right] \\
&= 1 - \mathop{\mathbb{E}}_{\mathsf{X}'\sim Q}\left[-\int_{-\infty}^{\infty} \ell_{-1}(s(\mathsf{X}') - t) \cdot \mathrm{TPR}'(t)\, dt\right] \\
&= 1 - \mathop{\mathbb{E}}_{\mathsf{X}'\sim Q}\left[\int_{-\infty}^{\infty} \ell_{-1}(s(\mathsf{X}') - t) \cdot p_S(t)\, dt\right]
\end{aligned}$$

13. This restriction ensures that we can employ the integral substitution formula in the proof.

$$= 1 - \mathop{\mathbb{E}}_{\mathsf{X}' \sim Q} \left[\int_{-\infty}^{\infty} \ell_{-1}(s(\mathsf{X}') - t) \cdot \mathop{\mathbb{E}}_{\mathsf{X} \sim P} \left[\delta_{s(\mathsf{X})}(t) \right] \, dt \right]$$
$$= 1 - \mathop{\mathbb{E}}_{\mathsf{X}' \sim Q, \mathsf{X} \sim P} \left[\ell_{-1}(s(\mathsf{X}') - s(\mathsf{X})) \right].$$

Thus,

$$\mathrm{AUC}(s; D, \ell) = \frac{1}{2} \int_0^1 (\mathrm{TPR}_\ell(\mathrm{FPR}^{-1}(\alpha)) + \mathrm{TNR}_\ell(\mathrm{TPR}^{-1}(\alpha))) \, d\alpha.$$

■

For a symmetric loss, we can thus interpret the ℓ-AUC as being the area under a curve that plots the false positive rates of a scorer against the ℓ-true positive rates, or equivalently, the true positive rates against the ℓ-true negative rates. (These two curves are not equivalent in general, but the area under them is.) Clearly, when $\ell = \ell_{01}$, the ℓ-power is the standard power (Equation 21), and the above is exactly the standard AUC.

5.4.2 BASIC PROPERTIES OF THE ℓ-AUC

Like the standard AUC, the ℓ-AUC does not depend on the base rate π. However, the ℓ-AUC does not inherit all properties of the standard AUC. First, in general the ℓ-AUC lies in $\mathbb{R}$, not necessarily $[0, 1]$. Second, the ℓ-AUC is not invariant to strictly monotone increasing transforms of the scoring functions. Indeed, the ℓ-AUC penalises the *magnitude* of differences between predictions. Intuitively, this makes the ℓ-AUC closer to a classification or class-probability estimation risk, as it is not sufficient to simply order instances well. Similar magnitude-sensitive metrics have been explored by Wu and Flach (2005); Ferri et al. (2005); for example, Wu and Flach (2005, Equation 5) proposed the "scored AUC",

$$\mathrm{AUC}_{\mathrm{scr}}(s; D) \doteq \mathop{\mathbb{E}}_{\mathsf{X} \sim P, \mathsf{X}' \sim Q} \left[\max(0, s(\mathsf{X}) - s(\mathsf{X}')) \right].$$

corresponding to the ℓ-AUC with non-convex loss $\ell_1(v) = (1 - v) \wedge 1$.

5.5 The ℓ-AUC and bipartite ranking risk

As mentioned, the AUC is a canonical measure of performance for bipartite ranking problems. Thus far, we have used the bipartite risk (Equation 16) as our measure of performance. In fact, these measures are equivalent: from the definition of the ℓ-AUC for a scorer s (Definition 19), it is apparent that it is a linear transformation of the bipartite ranking risk for the pair-scorer $\mathrm{Diff}(s)$. Equivalently, by Lemma 2, the ℓ-AUC may be seen as a linear transformation of the pairwise ranking risk over D_{BR}.

Lemma 22 *For any $D \in \Delta_{\mathcal{X} \times \{\pm 1\}}$, loss ℓ, and scorer $s : \mathcal{X} \to \mathbb{R}$,*

$$\mathrm{AUC}(s; D, \ell) = 1 - \mathbb{L}_{\mathrm{BR}}(s; D, \ell) \tag{32}$$
$$= 1 - \mathbb{L}(\mathrm{Diff}(s); D_{\mathrm{BR}}, \ell), \tag{33}$$

so that the ℓ-AUC can be seen as a linear transform of ℓ-bipartite ranking risk.

We may further relate the Bayes-optimal scorers for AUC and bipartite risk. For a distribution $D \in \Delta_{\mathcal{X} \times \{\pm 1\}}$, define the *Bayes-optimal ℓ-AUC* to be the supremal ℓ-AUC:

$$\mathrm{AUC}^*(D, \ell) = \sup_{s : \mathcal{X} \to \mathbb{R}} \mathrm{AUC}(s; D, \ell).$$

This supremal AUC is simply one minus the Bayes-optimal bipartite ranking risk; thus, it is a measure of the inherent difficulty of bipartite ranking with a given distribution D.

Representation of $\mathrm{AUC}(s; D)$	**Interpretation**	**Reference**
$\int_0^1 \mathrm{TPR}(\mathrm{FPR}^{-1}(\alpha))\, d\alpha$	Area under ROC curve	Equation 27
$\mathbb{P}_{X\sim P, X'\sim Q}[s(X) > s(X')] + \frac{1}{2} \cdot \mathbb{P}_{X\sim P, X'\sim Q}[s(X) = s(X')]$	Probability of random positive scoring higher than random negative	Equation 28
$1 - \mathop{\mathbb{E}}_{X\sim P, X'\sim Q}\left[\ell_{01}(1, s(X) - s(X'))\right]$	Average 0-1 accuracy on pairs	Equation 29
$1 - \mathbb{L}_{\mathrm{BR}}(\mathrm{Diff}(s); D, \ell_{01})$	One minus bipartite ranking risk	Equation 32
$1 - \frac{1}{2\pi(1-\pi)} \mathop{\mathbb{E}}_{(X,Y)\sim D}\left[\int_0^1 \mu_{\mathrm{Cal}(\cdot; D, s)}(c) \cdot \lambda_{\mathrm{CS}(c)}(Y, \mathrm{Cal}(X; D, s))\, dc\right]$	Weighted combination of cost-sensitive losses	Equation 34
$2 \cdot \mathop{\mathbb{E}}_{X\sim P}[\mathrm{BACC}(s(X); D, s)] - \frac{1}{2}$	Average balanced accuracy	Equation 39
$\mathop{\mathbb{E}}_{X\sim P}[\mathrm{TNR}(s(X))]$	Average rank of positive examples	Equation 37
$\mathop{\mathbb{E}}_{X'\sim Q}\left[\mathrm{TPR}(s(X'))\right]$	Average rank of negative examples	Equation 38

Table 7: Various representations for the AUC of a scorer $s : \mathcal{X} \to \mathbb{R}$ with respect to a distribution $D \in \Delta_{\mathcal{X}\times\{\pm 1\}}$. Each gives a different interpretation, and possibly means of estimating the AUC from samples.

Corollary 23 *For any $D \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ and loss ℓ,*

$$\mathrm{AUC}^*(D, \ell) = 1 - \mathbb{L}^*_{\mathrm{BR}}(D, \ell).$$

Proof Take the supremum of both terms in Equation 32. ■

5.6 Alternate representations of the AUC

We now outline several equivalent representations of the AUC, summarised in Figure 7. Each gives a different perspective about how it measures the performance of a scorer, as well as potentially different means of estimating it from samples. Many of these representations are specific to ℓ_{01}, but we begin with two related representations that hold for general ℓ.

5.6.1 THE SHUFORD REPRESENTATION

Suppose the loss ℓ is proper composite with link Ψ. Recall Shuford's integral representation (Equation 10),

$$\ell(y, v) = \int_0^1 w(c) \cdot \lambda_{\mathrm{CS}(c)}(y, \Psi^{-1}(v))\, dc.$$

We can apply this to the definition of pairwise ranking risk (Equation 19) to get a representation of the bipartite ranking risk in terms of cost-sensitive bipartite ranking risks, assuming Ψ has some symmetry.

Proposition 24 *For any $D \in \Delta_{\mathcal{X}\times\{\pm 1\}}$, $\ell \in \mathcal{L}_{\mathrm{SPC}}(\Psi)$ where $\Psi^{-1}(-v) = 1 - \Psi^{-1}(v)$ and scorer $s : \mathcal{X} \to \mathbb{R}$,*

$$\mathbb{L}_{\mathrm{BR}}(s; D, \ell) = \int_0^1 w(c) \cdot \mathbb{L}(\Psi^{-1} \circ \mathrm{Diff}(s); D_{\mathrm{BR}}, \lambda_{\mathrm{symm,CS}(c)})\, dc,$$

where $w(\cdot)$ is the weight function corresponding to the proper loss $\lambda = \ell \circ \Psi^{-1}$, and

$$(\forall c \in [0,1])\,(\forall u \in [0,1])\ \lambda_{\mathrm{symm,CS}(c)}(u) = \frac{\lambda_{\mathrm{CS}(c)}(+1, u) + \lambda_{\mathrm{CS}(c)}(-1, 1-u)}{2}.$$

Proof By the equivalence of bipartite ranking and classification on pairs, the given condition on the link function, and applying Shuford's representation to the partial losses $\ell_{\pm 1}$,

$$\begin{aligned}
\mathbb{L}_{\text{BR}}(s; D, \ell) &= \mathbb{L}(\text{Diff}(s); D_{\text{BR}}, \ell) \\
&= \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{X}')\sim(P\times Q)} \left[\ell_{\text{symm}}((\text{Diff}(s))(\mathsf{X},\mathsf{X}'))\right] \\
&= \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{X}')\sim(P\times Q)} \left[\frac{\ell_1((\text{Diff}(s))(\mathsf{X},\mathsf{X}')) + \ell_{-1}(-(\text{Diff}(s))(\mathsf{X},\mathsf{X}'))}{2}\right] \\
&= \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{X}')\sim(P\times Q)} \left[\frac{\lambda_1(\Psi^{-1}((\text{Diff}(s))(\mathsf{X},\mathsf{X}'))) + \lambda_{-1}(\Psi^{-1}(-(\text{Diff}(s))(\mathsf{X},\mathsf{X}')))}{2}\right] \\
&= \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{X}')\sim(P\times Q)} \left[\frac{\lambda_1(\Psi^{-1}((\text{Diff}(s))(\mathsf{X},\mathsf{X}'))) + \lambda_{-1}(1-\Psi^{-1}((\text{Diff}(s))(\mathsf{X},\mathsf{X}')))}{2}\right] \\
&= \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{X}')\sim(P\times Q)} \left[\int_0^1 w(c)\cdot\frac{\lambda_{\text{CS}(c)}(1,\Psi^{-1}((\text{Diff}(s))(\mathsf{X},\mathsf{X}'))) + \lambda_{\text{CS}(c)}(-1,1-\Psi^{-1}((\text{Diff}(s))(\mathsf{X},\mathsf{X}')))}{2}\right] \\
&= \int_0^1 w(c)\cdot\mathbb{L}(\Psi^{-1}\circ\text{Diff}(s); D_{\text{BR}}, \lambda_{\text{symm,CS}(c)})\,dc,
\end{aligned}$$

where

$$\lambda_{\text{symm,CS}(c)}(u) = \frac{\lambda_{\text{CS}(c)}(+1,u) + \lambda_{\text{CS}(c)}(-1,1-u)}{2}.$$

■

Without the assumption on Ψ above, one can still obtain an integral representation, but it will not cleanly relate to weighted combination of probability estimation loss risks.

Given the connection between the ℓ-AUC and bipartite ranking risk (Lemma 22), one might hope for a result of the form

$$\text{AUC}(s; D, \ell) = \int_0^1 w(c)\cdot\text{AUC}(s; D, \lambda_{\text{CS}(c)})\,dc.$$

However, the AUC can only be equated with the pairwise ranking risk when the latter uses a decomposable scorer (Equation 33). In the above equation, if we require $\Psi^{-1}\circ\text{Diff}(s)$ to be decomposable, then it must be that Ψ^{-1} is the identity function, i.e. the loss ℓ must be proper, and not merely proper composite. But recall that the ℓ-AUC is not defined for a proper loss, because its prediction space is not closed under negation. Thus, an integral representation cannot be found here, either.

Interestingly, if we allow the weights to depend on the scorer s, it *is* possible to get an expression for the AUC with strong resemblance to Shuford's result for proper losses, as we now explore.

5.6.2 HAND'S REPRESENTATION

We now generalise a result of Hand (2009, Section 4) to the case of general ℓ-AUC. Informally, the result is a representation of the ℓ-AUC as a weighted combination of cost-sensitive risks, where the weighting is distribution and scorer dependent. More specifically, the result states that the AUC of a scorer s is a weighted combination of cost-sensitive risks, where the weighting factor on costs depends on s, and the threshold for cost c is set to the optimal choice (c.f. Proposition 11) of $\text{Prb}^{-1}(c)$.

Proposition 25 *Given any $D \in \Delta_{\mathcal{X}\times\{\pm 1\}}$, loss ℓ, and scorer $s : \mathcal{X} \to \mathbb{R}$ such that* $\text{ROC}(s; D)$ *is differentiable and* $\text{Prb}(; D, s)$ *is differentiable and invertible,*

$$\text{AUC}(s; D, \ell) = 1 - \frac{1}{2\pi(1-\pi)}\int_0^1 V_S(c)\cdot\mathbb{L}(s; D, \ell_{\text{trans}(c)})\,dc$$

where the transformed loss $\ell_{\text{trans}(c)}$ with parameter c is

$$(\ell_{\text{trans}(c)})(y, v) = (\text{T}(\ell, c, \text{Prb}^{-1}(c)))(y, v)$$

$$(\mathrm{T}(\ell, c, t))(y, v) = (1 - c) \cdot [\![y = 1]\!] \cdot \ell_1(v - t) + c \cdot [\![y = -1]\!] \cdot \ell_{-1}(v - t),$$

and the weighting factor is

$$V_S(c) = (\mathrm{Prb}^{-1})'(c) \cdot \mu_S(\mathrm{Prb}^{-1}(c))$$

for marginal distribution over scores μ_S.

Proof By the rate-based representation from Equations 30, 31,

$$1 - \mathrm{AUC}(s; D, \ell) = \frac{1}{2} \int_{-\infty}^{\infty} (\mathrm{TNR}'(t) \cdot \mathrm{FNR}_\ell(t) + \mathrm{FNR}'(t) \cdot \mathrm{FPR}_\ell(t))\, dt.$$

Recall from §2.4 that we refer to the marginal density of scores by μ_S, and the class-conditional densities by p_S, q_S. Recall from Equation 22 that for any $t \in \mathbb{R}$,

$$\begin{aligned} \mathrm{TNR}'(t) &= q_S(t) \\ &= \frac{1}{1 - \pi} \cdot \mu_S(t) \cdot (1 - \mathrm{Prb}(t)), \end{aligned}$$

and similarly,

$$\mathrm{FNR}'(t) = \frac{1}{\pi} \cdot \mu_S(t) \cdot \mathrm{Prb}(t).$$

Thus, the integrand is

$$\begin{aligned} \iota(t) &= \mathrm{TNR}'(t) \cdot \mathrm{FNR}_\ell(t) + \mathrm{FNR}'(t) \cdot \mathrm{FPR}_\ell(t) \\ &= \frac{1}{\pi(1 - \pi)} \cdot \mu_S(t) \cdot \big(\pi \cdot (1 - \mathrm{Prb}(t)) \cdot \mathrm{FNR}_\ell(t) + (1 - \pi) \cdot \mathrm{Prb}(t) \cdot \mathrm{FPR}_\ell(t)\big) \\ &= \frac{1}{\pi(1 - \pi)} \cdot \mu_S(t) \cdot \mathbb{L}(s; D, \ell_{\mathrm{trans}(\mathrm{Prb}(t))}), \end{aligned}$$

where

$$\begin{aligned} (\ell_{\mathrm{trans}(c)})(y, v) &= (\mathrm{T}(\ell, c, \mathrm{Prb}^{-1}(c)))(y, v) \\ (\mathrm{T}(\ell, c, t))(y, v) &= (1 - c) \cdot [\![y = 1]\!] \cdot \ell_1(v - t) + c \cdot [\![y = -1]\!] \cdot \ell_{-1}(v - t), \end{aligned}$$

transforms the loss ℓ to use a cost weighting c and corresponding optimal threshold $\mathrm{Prb}^{-1}(c)$ (c.f. Proposition 11). Thus, returning the original integral,

$$\begin{aligned} 1 - \mathrm{AUC}(s; D, \ell) &= \frac{1}{2\pi(1 - \pi)} \int_{-\infty}^{\infty} \mu_S(t) \cdot \mathbb{L}(s; D, \ell_{\mathrm{trans}(\mathrm{Prb}(t))})\, dt \\ &= \frac{1}{2\pi(1 - \pi)} \int_0^1 V_S(c) \cdot \mathbb{L}(s; D, c)\, dc, \end{aligned}$$

using the substitution $c = \mathrm{Prb}(t)$, with $V_S(c) \doteq (\mathrm{Prb}^{-1})'(c) \cdot \mu_S(\mathrm{Prb}^{-1}(c))$. ■

The right hand side above features two seemingly opaque objects: a transformed version of the loss and a weighting factor, both of which depend on the score-to-probability transformation $\mathrm{Prb}(\cdot\,; D, s)$. Fortunately, both of these simplify when we consider ℓ_{01} (corresponding to the standard AUC), and calibrated scorers. First, it is easy to check that for ℓ_{01}, the transformed loss is the cost-sensitive loss with threshold t,

$$(\mathrm{T}(\ell, c, t))(y, v) = \ell_{\mathrm{CS}(c,t)}(y, v)$$

where $\ell_{\mathrm{CS}(c,t)}$ is as in Equation 26. Further, if we calibrate our scorer, we find

$$(\mathrm{Prb} \circ \mathrm{Cal})(t) = \mathbb{P}[\mathsf{Y} = 1 | s(\mathsf{X}) = \mathrm{Prb}^{-1}(t)] = t$$

$$V_{\mathrm{Cal}(\cdot;D,s)}(c) = \mu_C(c),$$

where C is the distribution of the calibrated scorer $\mathrm{Cal}(\cdot\,; D, s)$. When the calibration transform is invertible, we may use this to express the standard AUC as the following, which is also a consequence of Hand (2009, Equation 6).

Corollary 26 *Given any $D \in \Delta_{\mathcal{X}\times\{\pm 1\}}$, loss ℓ, and scorer $s : \mathcal{X} \to \mathbb{R}$ such that* $\mathrm{ROC}(s; D)$ *is differentiable and* $\mathrm{Cal}(\cdot\,; D, s)$ *is strictly monotone increasing,*

$$\mathrm{AUC}(s; D) = 1 - \frac{1}{2\pi(1-\pi)} \cdot \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim D} \left[\int_0^1 \mu_C(c) \cdot \lambda_{\mathrm{CS}(c)}(\mathsf{Y}, \mathrm{Cal}(\mathsf{X}; D, s))\, dc \right]. \tag{34}$$

If in particular s is calibrated with respect to D,

$$\mathrm{AUC}(s; D) = 1 - \frac{1}{2\pi(1-\pi)} \cdot \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim D} \left[\int_0^1 \mu_S(c) \cdot \lambda_{\mathrm{CS}(c)}(\mathsf{Y}, s(\mathsf{X}))\, dc \right].$$

Equation 34 gives two interesting perspectives on the AUC. First, when it is possible to calibrate a scorer without losing information, the AUC can be thought of as implicitly calibrating a scorer before computing a particular risk. Second, by Shuford's representation (Equation 10), the risk computed is in fact identical to that for a proper loss, with the caveat that one considers a *score- and distribution-dependent* weight function[14] $w(c) = \mu_C(c)$. Thus, the AUC is equivalent to the risk of a score- and distribution-dependent proper loss. (This is the finding of, for example, Hernández-Orallo et al. (2012, Theorem 34), which equates the AUC to the squared loss risk under a special case. We illustrate this equivalence empirically in Appendix K.) In particular, for a fixed distribution D, the AUC employs a different weighting for different scorers. Consequently, Hand (2009) calls the AUC an "incoherent" measure of classifier performance. Hand (2009, Section 6) proposes replacing this scorer dependent weight with one derived from the Beta family:

$$w(c; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \cdot c^{\alpha-1} \cdot (1-c)^{\beta-1},$$

where $B(\alpha, \beta)$ is the normaliser for the Beta distribution. From Shuford's integral representation, it is apparent that the corresponding risk exactly corresponds to that of a proper composite loss. Indeed, the Beta family was proposed as a template for generating a proper loss in Buja et al. (2005, Section 11). For another perspective on this issue from the perspective of threshold selection, see Flach et al. (2011); Hernández-Orallo et al. (2012).

Another perspective on the "incoherence" can be gained from Proposition 13. Suppose we have (calibrated) scorers s_1, s_2 such that $\mathrm{AUC}(s_1; D) > \mathrm{AUC}(s_2; D)$. If it is further true that the ROC curve of s_1 dominates that of s_2, then we know that s_1 will have lower risk than s_2 with respect to any proper composite risk, or equivalently any (distribution-independent) weighted combination of cost-sensitive risks. Thus, it is "coherent" to compare the two scorers based on their AUC in this case, because every other measure will result in s_1 being adjudged superior. When the ROC curves of the two scorers cross, however, the AUC is an incomplete measure of performance; with some proper losses, s_1 will be favoured over s_2, and vice versa.

5.6.3 RATE-BASED REPRESENTATIONS

We proceed with some rate-based representations for the AUC. From a graphical perspective, these all derive from the fact that the AUC is the area under the ROC curve, and that this area is invariant to rotation of the horizontal and vertical axes i.e. instead of plotting the false positive versus true positive rate, we can equally plot the true negative versus false negative rate.

14. By definition $\int_0^1 \mu_S(c)\, dc = 1$, and so this equivalence can only be established to proper losses that satisfy $\int_0^1 w(c)\, dc < \infty$. This rules out, for example, logistic and exponential risk being equivalent to AUC.

Proposition 27 *Given any $D \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ and scorer $s : \mathcal{X} \to \mathbb{R}$ with differentiable ROC curve and invertible false- and true-positive rates,*

$$\begin{aligned}\mathrm{AUC}(s; D) &= \int_0^1 \mathrm{TPR}(\mathrm{FPR}^{-1}(\alpha))\, d\alpha \\ &= \int_0^1 \mathrm{TPR}(\mathrm{TNR}^{-1}(\alpha))\, d\alpha \\ &= \int_0^1 \mathrm{TNR}(\mathrm{TPR}^{-1}(\alpha))\, d\alpha \\ &= \int_0^1 \mathrm{TNR}(\mathrm{FNR}^{-1}(\alpha))\, d\alpha.\end{aligned}$$

Proof The first equation is simply Equation 27. The subsequent expressions follow from a few simple facts. First, if $f(x) = 1 - g(x)$ and f is invertible with inverse f^{-1}, then

$$g^{-1}(x) = f^{-1}(1 - x).$$

This implies that

$$\begin{aligned}\mathrm{FPR}^{-1}(\alpha) &= \mathrm{TNR}^{-1}(1 - \alpha) \\ \mathrm{TPR}^{-1}(\alpha) &= \mathrm{FNR}^{-1}(1 - \alpha).\end{aligned}$$

Second, for any f,

$$\int_0^1 f(1 - x)\, dx = \int_0^1 f(x)\, dx.$$

Combined with the above, this implies

$$\begin{aligned}\int_0^1 \mathrm{TPR}(\mathrm{FPR}^{-1}(\alpha))\, d\alpha &= \int_0^1 \mathrm{TPR}(\mathrm{TNR}^{-1}(\alpha))\, d\alpha \\ \int_0^1 \mathrm{TNR}(\mathrm{TPR}^{-1}(\alpha))\, d\alpha &= \int_0^1 \mathrm{TNR}(\mathrm{FNR}^{-1}(\alpha))\, d\alpha.\end{aligned}$$

Third, for any f, g, integration by parts implies that

$$\int_a^b f'(x)g(x)\, dx = f(b)g(b) - f(a)g(a) - \int_a^b f(x)g'(x)\, dx. \tag{35}$$

Fourth, for any f, g such that $g(a) = 0, g(b) = 1$,

$$\int_0^1 f(g^{-1}(t))\, dt = \int_a^b g'(x)f(x)\, dx. \tag{36}$$

This implies that:

$$\begin{aligned}\int_0^1 \mathrm{TPR}(\mathrm{TNR}^{-1}(\alpha))\, d\alpha &= \int_{-\infty}^{\infty} \mathrm{TNR}'(t) \cdot \mathrm{TPR}(t)\, dt \\ &= -\int_{-\infty}^{\infty} \mathrm{TPR}'(t) \cdot \mathrm{TNR}(t)\, dt \text{ by Equation 35} \\ &= \int_0^1 \mathrm{TNR}(\mathrm{TPR}^{-1}(\alpha))\, d\alpha \text{ by Equation 36.}\end{aligned}$$

Now recalling the definition of the AUC (Definition 14) as $\int_0^1 \text{TPR}(\text{FPR}^{-1}(\alpha))\, d\alpha$, we see that we have proved the proposition. ■

From the above proof, we see that one can equivalently express the AUC as a weighted average of individual rates over a range of thresholds.

Corollary 28 *Given any $D \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ and scorer $s : \mathcal{X} \to \mathbb{R}$ with differentiable ROC curve and invertible false- and true-positive rates,*

$$\begin{aligned}\text{AUC}(s; D) &= -\int_{-\infty}^{\infty} \text{FPR}'(t) \cdot \text{TPR}(t)\, dt \\ &= \int_{-\infty}^{\infty} \text{TNR}'(t) \cdot \text{TPR}(t)\, dt \\ &= -\int_{-\infty}^{\infty} \text{TPR}'(t) \cdot \text{TNR}(t)\, dt \\ &= \int_{-\infty}^{\infty} \text{FNR}'(t) \cdot \text{TNR}(t)\, dt.\end{aligned}$$

The weighting over thresholds as expressed above is not particularly intuitive, but recall from Equation 22 that the derivatives of the rates are the corresponding class-conditional densities of the scores. That means that we can interpret the above choice as equivalently drawing thresholds from these distributions. This is explored in the next section.

5.6.4 RANK REPRESENTATION

We now show how the AUC can be interpreted as the average of *ranks* of the instances, where the average is over thresholds drawn in accordance with the distribution of scores.

Corollary 29 *Given any $D = \langle P, Q, \pi \rangle \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ and scorer $s : \mathcal{X} \to \mathbb{R}$,*

$$\text{AUC}(s; D) = \mathop{\mathbb{E}}_{\mathsf{X}\sim P} \left[\text{TNR}(s(\mathsf{X}))\right] \tag{37}$$

$$= \mathop{\mathbb{E}}_{\mathsf{X}'\sim Q} \left[\text{TPR}(s(\mathsf{X}'))\right]. \tag{38}$$

Proof This follows immediately from Corollary 28 and the definition of the derivatives of the rates. Alternately, by Proposition 17, and rewriting probabilities as expectations,

$$\begin{aligned}\text{AUC}(s; D) &= \mathbb{P}_{\mathsf{X}\sim P, \mathsf{X}'\sim Q}[s(\mathsf{X}) > s(\mathsf{X}')] + \frac{1}{2} \cdot \mathbb{P}_{\mathsf{X}\sim P, \mathsf{X}'\sim Q}[s(\mathsf{X}) = s(\mathsf{X}')] \\ &= \mathop{\mathbb{E}}_{\mathsf{X}\sim P} \left[\mathop{\mathbb{E}}_{\mathsf{X}'\sim Q} \left[[\![s(\mathsf{X}) > s(\mathsf{X}')]\!] + \frac{1}{2} [\![s(\mathsf{X}) = s(\mathsf{X}')]\!] \right] \right] \\ &= \mathop{\mathbb{E}}_{\mathsf{X}\sim P} \left[\text{TNR}(s(\mathsf{X}))\right].\end{aligned}$$

Swapping the order of expectations in the other direction gives Equation 38. ■

On a finite training set, the empirical version of $\text{TNR}(s(x))$ is related to the (normalised version of) what is typically called the "rank" of an instance $x \in \mathcal{X}$, where a higher rank is better. Specifically, the empirical $\text{TNR}(s(x))$ counts the fraction of negative instances that x is scored higher than. In this sense, the AUC can be seen as measuring the average rank of the positive examples.

5.6.5 BALANCED ACCURACY REPRESENTATION

Our final representation of the AUC more explicitly relates it to a measure of classification performance: we show how to rewrite it as the average "balanced accuracy" across a range of thresholds, where the balanced accuracy is the average of the accuracies on the positive and negative class individually (Chan and Stolfo, 1998),

$$\mathrm{BACC}(t; D, s) \doteq \frac{\mathrm{TPR}(t; D, s) + \mathrm{TNR}(t; D, s)}{2}.$$

This suggests that the AUC explicitly considers good performance on both classes simultaneously, which indicates it is useful for problems with class imbalance (Ling and Li, 1998). For a related representation on a finite training set, see Flach et al. (2011, Theorem 4, 5), while for a different proof strategy, see Menon et al. (2015, Proposition 20).

Proposition 30 *For any $D \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ and scorer $s : \mathcal{X} \to \mathbb{R}$ with differentiable ROC curve and invertible false- and true-positive rates,*

$$\begin{aligned}\mathrm{AUC}(s; D) &= 2 \cdot \mathop{\mathbb{E}}_{\mathsf{X}\sim P}[\mathrm{BACC}(s(\mathsf{X}); D, s)] - \frac{1}{2} \\ &= 2 \cdot \mathop{\mathbb{E}}_{\mathsf{X}'\sim Q}\left[\mathrm{BACC}(s(\mathsf{X}'); D, s)\right] - \frac{1}{2}.\end{aligned} \tag{39}$$

Proof This is from Corollary 29 and the fact that

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathsf{X}\sim P}[\mathrm{TPR}(s(\mathsf{X}); D, s)] &= \mathop{\mathbb{E}}_{\mathsf{S}\sim P_S}[\mathrm{TPR}(\mathsf{S}; D, s)] \\ &= \int_{-\infty}^{\infty} -\mathrm{TPR}'(t) \cdot \mathrm{TPR}(t)\, dt \\ &= \int_0^1 u\, du \text{ with } u = \mathrm{TPR}(t) \\ &= \frac{1}{2}.\end{aligned} \tag{40}$$

A similar argument yields the second identity. ∎

The representation of Equation 39 can be contrasted with the risk for a proper composite loss. Using Shuford's formula (Equation 10), for a proper composite loss ℓ with surjective and differentiable link Ψ, we have (Reid and Williamson, 2011, Proposition 20)

$$\begin{aligned}\mathbb{L}(s; D, \ell) &= \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim D}\left[\lambda(\mathsf{Y}, \Psi^{-1}(s(\mathsf{X})))\right] \\ &= \int_0^1 w(c) \cdot \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim D}\left[\lambda_{\mathrm{CS}(c)}(\mathsf{Y}, \Psi^{-1}(s(\mathsf{X})))\right] \\ &= \int_0^1 w(c) \cdot ((1-\pi) \cdot c \cdot \mathrm{FPR}(\Psi(c); D, s) + \pi \cdot (1-c) \cdot \mathrm{FNR}(\Psi(c); D, s))\, dc \\ &= \int_{-\infty}^{\infty} \frac{w(\Psi^{-1}(t))}{\Psi'(\Psi^{-1}(t))} \cdot ((1-\pi) \cdot \Psi^{-1}(t) \cdot \mathrm{FPR}(t; D, s) + \pi \cdot (1 - \Psi^{-1}(t)) \cdot \mathrm{FNR}(t; D, s))\, dt.\end{aligned}$$

Compared to Equation 39, we see that the proper composite risk has potentially asymmetric, but distribution independent, weights on the FPR and FNR. We also observe that the weights on the FPR and FNR are not equal, and vary with the thresholds. As per Hand's representation (Equation 34), we may find that for a fixed distribution, there is a choice of link Ψ and weight w such that the two representations agree.

5.7 Relation to existing work

Lemma 22, which relates the AUC with the pairwise ranking (and hence pairwise classification) risk, is well known for the case of 0-1 loss (Kotlowski et al., 2011; Agarwal, 2014). The extension to an arbitrary loss ℓ, while simple, is to our knowledge new. More generally, our definition of the ℓ-AUC as a generalisation of the standard AUC appears to be new, although in the special case where ℓ is a convex margin loss, the risk counterpart has been discussed (Clémençon et al., 2008). The study of integral representations of the ℓ-AUC is to our knowledge new.

The representations in §5.6 are not new, although several of them only appear to have been stated for a finite training set.

6. Relating the Bayes risk and regret to divergences

The previous sections studied the bipartite risk for an arbitrary scorer. In this section, we study the bipartite risk for the *Bayes-optimal* scorer, as well as the *regret* or excess risk for an arbitrary scorer. These help understand the inherent difficulty of a bipartite ranking problem, and formalise the sense in which "closeness" to the optimal scorer relates to the minimisation of the risk. Our characterisations rely on two classes of divergences between distributions, namely, the f- and Bregman-divergences. A review of the role of these divergences in characterising Bayes-risk and regret for classification is provided in Appendix E.

6.1 Warm-up: Bayes-optimal pairwise ranking risk and regret

As pairwise ranking is readily shown to be equivalent to binary classification over pairs of instances (see §10.4), plugging in a pairwise ranking distribution $R \in \Delta_{\mathcal{X}\times\mathcal{X}\times\{\pm 1\}}$ into existing results for classification immediately implies the following.

Proposition 31 *For any $R = \langle P_{\text{pair}}, Q_{\text{pair}}, \pi_{\text{pair}}\rangle \in \Delta_{\mathcal{X}\times\mathcal{X}\times\{\pm 1\}}$, convex $f : \mathbb{R}_+ \to \mathbb{R}$, and loss ℓ with conditional Bayes risk*

$$(\forall \eta \in (0,1))\; L^*(\eta;\ell) = -\frac{1-\eta}{1-\pi_{\text{pair}}} \cdot f\left(\frac{1-\pi_{\text{pair}}}{\pi_{\text{pair}}} \cdot \frac{\eta}{1-\eta}\right),$$

the Bayes pairwise ranking risk can be written

$$\mathbb{L}^*(R,\ell) = L^*(\pi_{\text{pair}};\ell) - \mathbb{I}_f(P_{\text{pair}}, Q_{\text{pair}}). \tag{41}$$

Conversely, Equation 41 holds for any $R = \langle P_{\text{pair}}, Q_{\text{pair}}, \pi_{\text{pair}}\rangle \in \Delta_{\mathcal{X}\times\mathcal{X}\times\{\pm 1\}}$, loss ℓ with concave conditional Bayes risk $L^* : [0,1] \to \mathbb{R}_+$, and $f : \mathbb{R}_+ \to \mathbb{R}$ defined by

$$(\forall t \in \mathbb{R})\; f(t) \doteq L^*(\pi_{\text{pair}};\ell) - (\pi_{\text{pair}} \cdot t + 1 - \pi_{\text{pair}}) \cdot L^*\left(\frac{\pi_{\text{pair}} \cdot t}{\pi_{\text{pair}} \cdot t + 1 - \pi_{\text{pair}}};\ell\right).$$

An important example of the above is for the case of the 0-1 loss ℓ_{01}, where the corresponding f-divergence is the variational divergence (or total variation distance) $V(\cdot,\cdot)$, given by

$$V(P,Q) \doteq \sup_{\mathcal{A}\subseteq\mathcal{X}} 2 \cdot |P(\mathcal{A}) - Q(\mathcal{A})| = \int_{\mathcal{X}} |p(x) - q(x)|\, dx.$$

Proposition 31 thus implies that the Bayes pairwise ranking risk is an affine transformation of $V(P_{\text{pair}}, Q_{\text{pair}})$.

We similarly have a simple expression for the pairwise ranking regret with a proper loss.

Proposition 32 *For any $R = \langle M_{\text{pair}}, \eta_{\text{Pair}}\rangle \in \Delta_{\mathcal{X}\times\mathcal{X}\times\{\pm 1\}}$, $\ell \in \mathcal{L}_{\text{SPC}}(\Psi)$, and pair-scorer $s_{\text{Pair}} : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$,*

$$\text{regret}(s_{\text{Pair}}; R, \ell) = \mathbb{B}_{-L^*}(\eta_{\text{Pair}}, \Psi^{-1} \circ s_{\text{Pair}})$$

where in an abuse of notation $L^* = L^*(\cdot;\ell)$.

We now see how these results can be translated to the bipartite ranking setting.

6.2 Bayes-optimal bipartite risk as an f-divergence

For bipartite ranking, we can hope to exploit the connection between bipartite and pairwise ranking (Lemma 2), and derive analogues of the above for the distribution D_{BR}. A subtlety is that, as noted in Proposition 3, it is *not* necessarily true that $\mathbb{L}^*_{\mathrm{BR}}(D, \ell) = \mathbb{L}^*(D_{\mathrm{BR}}, \ell)$. Therefore, to translate the previous results, we need an additional condition ensuring this holds, which is simply that $\ell \in \mathcal{L}_{\mathrm{Decomp}}$.

Proposition 33 *For any $D = \langle P, Q, \pi \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$, convex $f : \mathbb{R}_+ \to \mathbb{R}$, and loss $\ell \in \mathcal{L}_{\mathrm{Decomp}}$ with conditional Bayes risk*

$$(\forall \eta \in (0,1))\, L^*(\eta; \ell) = -2 \cdot (1-\eta) \cdot f\left(\frac{\eta}{1-\eta}\right),$$

the Bayes-risk can be written

$$\mathbb{L}^*_{\mathrm{BR}}(D, \ell) = L^*(1/2; \ell) - \mathbb{I}_f(P \times Q, Q \times P). \tag{42}$$

Conversely, Equation 42 holds for any $D = \langle P, Q, \pi \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$, loss $\ell \in \mathcal{L}_{\mathrm{Decomp}}$ with concave conditional Bayes risk $L^*(\cdot; \ell) : [0,1] \to \mathbb{R}_+$, and $f : \mathbb{R}_+ \to \mathbb{R}$ defined by

$$(\forall t \in \mathbb{R})\, f(t) \doteq L^*(1/2; \ell) - \frac{1+t}{2} \cdot L^*\left(\frac{t}{1+t}; \ell\right).$$

Proof By Proposition 3, the assumption on ℓ inducing a decomposable Bayes-optimal pair-scorer for D_{BR} implies we can equate $\mathbb{L}^*_{\mathrm{BR}}(D, \ell)$ and $\mathbb{L}^*(D_{\mathrm{BR}}, \ell)$. We then apply Proposition 31 to D_{BR}, so that

$$\begin{aligned}
\mathbb{L}^*_{\mathrm{BR}}(D, \ell) &= \mathbb{L}^*(D_{\mathrm{BR}}, \ell) \text{ by Proposition 3} \\
&= L^*(\pi_{\mathrm{pair}}; \ell) - \mathbb{I}_f(P_{\mathrm{pair}}, Q_{\mathrm{pair}}) \text{ by Proposition 31} \\
&= L^*(1/2; \ell) - \mathbb{I}_f(P \times Q, Q \times P) \text{ by Appendix B .}
\end{aligned}$$

The other direction follows similarly. ∎

As we shall see in Proposition 44, the requirement that $\ell \in \mathcal{L}_{\mathrm{Decomp}}$ for a proper composite loss is equivalent to a condition on its link function. Importantly, there is *no* restriction on the underlying proper loss itself. Therefore, the above holds for a large class of losses. One such example is the 0-1 loss $\ell = \ell_{01}$, where we have following relationship between the Bayes-optimal AUC and the variational divergence between the product measures $P \times Q$ and $Q \times P$.

Corollary 34 *Given any $D = \langle P, Q, \pi \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$, the Bayes-optimal bipartite ranking risk is related to the variational divergence between the product distributions $P \times Q$ and $Q \times P$ via:*

$$\mathbb{L}^*(D_{\mathrm{BR}}, \ell_{01}) = \frac{1}{2} - \frac{1}{4} \cdot V(P \times Q, Q \times P).$$

Proof This follows from Proposition 33 and the fact that for ℓ_{01},

$$(\forall \eta \in [0,1])\, L^*(\eta; \ell_{01}) = \eta \wedge (1-\eta) = \frac{1}{2} - \left|\eta - \frac{1}{2}\right|,$$

with $L^*(1/2; \ell_{01}) = 1/2$. It is easy to check that this corresponds to $f(t) = |t-1|/4 + (1-t)/4$, which is a scaled version of the convex generator for the variational divergence. ∎

By Corollary 23 and Equation 48, Corollary 34 is equivalent to a result of Torgersen (1991, pg. 582),

$$\mathrm{AUC}^*(D) = \frac{1}{2} + \frac{1}{4} \cdot V(P \times Q, Q \times P).$$

This may be further manipulated to explicitly express the Bayes-optimal AUC in terms of the concentration of the values of η (Clémençon et al., 2008),

$$\text{AUC}^*(D) = \frac{1}{2} + \frac{1}{4\pi(1-\pi)} \cdot \mathop{\mathbb{E}}_{\mathsf{X} \sim M, \mathsf{X}' \sim M} \left[\left| \eta(\mathsf{X}) - \eta(\mathsf{X}') \right| \right].$$

This expression may be further related to the earth mover's distance (or L_1-Wasserstein metric) between the class-conditional distribution of scores (Clémençon et al., 2009).

6.3 Bipartite ranking regret as a generative Bregman divergence

The bipartite ranking regret for proper composite losses may similarly be re-expressed by exploiting the reduction to classification on pairs. We again need to restrict ourselves to those proper composite losses that induce a decomposable Bayes-optimal pair-scorer.

Proposition 35 *Pick any $D \in \Delta_{\mathcal{X} \times \{\pm 1\}}$ with derived pairwise ranking distribution $D_{\text{BR}} = \langle M_{\text{pair}}, \eta_{\text{Pair}} \rangle$, and any $\ell \in \mathcal{L}_{\text{SPC}}(\Psi) \cap \mathcal{L}_{\text{Decomp}}$. Then, for any scorer $s : \mathcal{X} \to \mathbb{R}$,*

$$\text{regret}_{\text{BR}}(s; D, \ell) = \mathbb{B}_{-L^*}(\eta_{\text{Pair}}, \Psi^{-1} \circ \text{Diff}(s))$$

where in an abuse of notation $L^* \doteq L^*(\cdot; \ell)$.

Proof By definition of the bipartite regret (Equation 18),

$$\begin{aligned}
\text{regret}_{\text{BR}}(s; D, \ell) &= \mathbb{L}_{\text{BR}}(s; D, \ell) - \mathbb{L}^*_{\text{BR}}(D, \ell) \\
&= \mathbb{L}(\text{Diff}(s); D_{\text{BR}}, \ell) - \mathbb{L}^*_{\text{BR}}(D, \ell) \\
&= \text{regret}(\text{Diff}(s); D_{\text{BR}}, \ell) + \mathbb{L}^*(D_{\text{BR}}, \ell) - \mathbb{L}^*_{\text{BR}}(D, \ell) \\
&= \mathbb{B}_{-L^*}(\eta_{\text{Pair}}, \Psi^{-1} \circ \text{Diff}(s)) + \mathbb{L}^*(D_{\text{BR}}, \ell) - \mathbb{L}^*_{\text{BR}}(D, \ell),
\end{aligned}$$

where the last line is by the standard expression for the regret with respect to a proper composite loss (Proposition 78). We thus need $\mathbb{L}^*(D_{\text{BR}}, \ell) = \mathbb{L}^*_{\text{BR}}(D, \ell)$, which by Proposition 3 is true iff $\ell \in \mathcal{L}_{\text{Decomp}}$. ■

In the case of $\ell = \ell_{01}$, the regret can be seen to measure the concentration of η in the region where the candidate scorer s disagrees with η, as is well known.

Corollary 36 ((Clémençon et al., 2008), (Agarwal, 2014, Theorem 11)) *For any $D = \langle M, \eta \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$ and scorer $s : \mathcal{X} \to \mathbb{R}$,*

$$\text{regret}_{\text{BR}}(s; D, \ell_{01}) = \mathop{\mathbb{E}}_{\mathsf{X} \sim M, \mathsf{X}' \sim M} \left[\left| \eta(\mathsf{X}) - \eta(\mathsf{X}') \right| \cdot \mathbb{I}(s, \eta; \mathsf{X}, \mathsf{X}') \right]$$

where

$$\mathbb{I}(s, \eta; \mathsf{X}, \mathsf{X}') = \left[\!\!\left[(\eta(\mathsf{X}) - \eta(\mathsf{X}')) \cdot (s(\mathsf{X}) - s(\mathsf{X}')) < 0 \right]\!\!\right] + \frac{1}{2} \cdot \left[\!\!\left[\eta(\mathsf{X}) = \eta(\mathsf{X}') \right]\!\!\right].$$

6.4 Relation to existing work

As noted above, the connection between Bayes risks and f-divergences in classification is well known (Österreicher and Vajda, 1993). For the bipartite ranking problem, the connection between the AUC and variational divergence has been made by (Torgersen, 1991; Reid and Williamson, 2011). The extension to the case of general ℓ-bipartite risk (Proposition 33) is simple in hindsight, as the variational divergence is well known to correspond to the use of 0-1 loss; however, to our knowledge, the extension is novel.

7. Bayes-optimal scorers for bipartite ranking

The previous section studied the risk associated with a Bayes-optimal scorer for bipartite ranking. We now characterise the set of Bayes-optimal scorers itself. Knowledge of the optimal scorers gives further insight into the problem, and helps relate it to the more familiar tasks of binary classification and class-probability estimation. As we shall see in the next section, this will also help in establishing the statistical consistency of the minimisation of surrogate losses on pairs for the task of AUC maximisation.

Before proceeding, we first recall the Bayes-optimal scorers for the latter problems given some $D = \langle P, Q, \pi \rangle = \langle M, \eta \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$. Table 8 summarises the findings of this section.

Loss type	Bayes-optimal (pair-)scorers	Reference
Classification-calibrated	$\mathcal{S}^*(D, \ell) \subseteq \left\{ s : \mathcal{X} \to \mathbb{R} : \begin{array}{r} \eta(x) \neq 1/2 \implies \\ \operatorname{sign}(s(x)) = \operatorname{sign}(2\eta(x) - 1) \end{array} \right\}$	Equation 43
	$\mathcal{S}^*(D_{\mathrm{BR}}, \ell) \subseteq \left\{ s_{\mathrm{Pair}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R} : \begin{array}{r} \eta(x) \neq \eta(x') \implies \\ \operatorname{sign}(s_{\mathrm{Pair}}(x, x')) = \\ \operatorname{sign}(\eta(x) - \eta(x')) \end{array} \right\}$	Equation 49
	$\{\phi \circ \eta\} \subseteq \mathcal{S}^*_{\mathrm{BR}}(D, \ell_{01}) = \{s : \mathcal{X} \to \mathbb{R} : \eta = \phi \circ s\}$	Proposition 42
Proper composite with link Ψ	$\{\Psi \circ \eta\} \subseteq \mathcal{S}^*(D, \ell)$	Equation 44
	$\{\Psi \circ \sigma \circ \mathrm{Diff}(\sigma^{-1} \circ \eta)\} \subseteq \mathcal{S}^*(D_{\mathrm{BR}}, \ell)$	Equation 50
	$\mathcal{S}^*_{\mathrm{BR}}(D, \ell) = \{\Psi \circ \eta + b : b \in \mathbb{R}\}$ for $\Psi \in \Sigma_{\mathrm{sig}}$	Corollary 45

Table 8: Bayes-optimal scorers and pair-scorers for various classification and bipartite ranking risks.

7.1 Binary classification

Consider a binary classification problem with distribution $D = \langle M, \eta \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$. If the loss ℓ is classification calibrated (Equation 7), every Bayes-optimal scorer must have the same sign as $\eta(x) - 1/2$, with the prediction for $\eta(x) = 1/2$ being irrelevant (Bartlett et al., 2006):

$$\mathcal{S}^*(D, \ell) \subseteq \{s : \mathcal{X} \to \mathbb{R} : \eta(x) \neq 1/2 \implies \operatorname{sign}(s(x)) = \operatorname{sign}(2\eta(x) - 1)\}\,. \tag{43}$$

When ℓ is the 0-1 loss, the above is an equality (Devroye et al., 1996). Thus, for ℓ_{01}, what is of interest is determining whether or not each instance has a greater than random chance of being labelled positive.

When ℓ is a proper composite loss with link Ψ, from the definition of properness (Equation 9) we can specify one minimiser of the conditional risk, which applied pointwise gives:

$$\{\Psi \circ \eta\} \subseteq \mathcal{S}^*(D, \ell). \tag{44}$$

This is an equality if and only if ℓ is strictly proper composite. Thus, a strictly proper composite loss requires precise information about η, unlike ℓ_{01}. Observe that $\Psi \circ \eta$ may be trivially transformed to give an optimal scorer for ℓ_{01}; thus, exactly solving class-probability estimation also solves binary classification. For an approximate solution, one can bound the excess ℓ_{01} error via a surrogate regret bound (Reid and Williamson, 2009).

7.2 Pairwise ranking

Recall from §3.4 that pairwise ranking is identical to binary classification over pairs of instances. Thus, in the pairwise ranking setting with distribution $R = \langle M_{\mathrm{pair}}, \eta_{\mathrm{Pair}} \rangle \in \Delta_{\mathcal{X} \times \mathcal{X} \times \{\pm 1\}}$, the above results can be

translated. For a classification calibrated loss, the Bayes-optimal pair-scorers must have the same sign as $\eta_{\text{Pair}}(x, x') - 1/2$, with the prediction for $\eta_{\text{Pair}}(x, x') = 1/2$ being irrelevant:

$$\mathcal{S}^*(R, \ell) \subseteq \left\{ s_{\text{Pair}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R} : \eta_{\text{Pair}}(x, x') \neq 1/2 \implies \text{sign}(s_{\text{Pair}}(x, x')) = \text{sign}(2\eta_{\text{Pair}}(x, x') - 1) \right\}. \tag{45}$$

When ℓ is the 0-1 loss, the above is an equality.

When ℓ is a proper composite loss with link Ψ, by definition we can specify one Bayes-optimal pair-scorer, though there may be others:

$$\{\Psi \circ \eta_{\text{Pair}}\} \subseteq \mathcal{S}^*(R, \ell). \tag{46}$$

This is an equality if and only if ℓ is strictly proper composite.

7.3 Bipartite ranking

The relationship between bipartite and pairwise ranking (Lemma 2) suggests we can simply compute the Bayes-optimal scorers for pairwise ranking with D_{BR}. Specifically, let $D_{\text{BR}} = \langle M_{\text{pair}}, \eta_{\text{Pair}} \rangle$. To determine $\mathcal{S}^*(D_{\text{BR}}, \ell)$, we need the following elementary but important property of η_{Pair}.

Lemma 37 *For any $D = \langle M, \eta \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$, D_{BR} has observation-conditional distribution given by*

$$\eta_{\text{Pair}} = \sigma \circ \text{Diff}(\sigma^{-1} \circ \eta), \tag{47}$$

where $\sigma(\cdot)$ denotes the sigmoid function (Equation 2).

Proof Suppose $(\mathsf{X}, \mathsf{X}', \mathsf{Z}) \sim D_{\text{BR}}$. Recall $\mathbb{P}[\mathsf{Z} = +1] = \frac{1}{2}$. Then,

$$\begin{aligned}
(\forall x, x' \in \mathcal{X})\, \eta_{\text{Pair}}(x, x') &= \mathbb{P}[\mathsf{Z} = +1 | \mathsf{X} = x, \mathsf{X}' = x'] \\
&= \frac{\mathbb{P}[\mathsf{X} = x, \mathsf{X}' = x' | \mathsf{Z} = +1] \cdot \mathbb{P}[\mathsf{Z} = +1]}{\mathbb{P}[\mathsf{X} = x, \mathsf{X}' = x']} \\
&= \frac{\mathbb{P}[\mathsf{X} = x | \mathsf{Z} = +1] \cdot \mathbb{P}[\mathsf{X}' = x' | \mathsf{Z} = +1] \cdot \mathbb{P}[\mathsf{Z} = +1]}{\mathbb{P}[\mathsf{X} = x, \mathsf{X}' = x']} \\
&= \frac{\mathbb{P}[\mathsf{X} = x | \mathsf{Z} = +1] \cdot \mathbb{P}[\mathsf{X}' = x' | \mathsf{Z} = +1]}{\mathbb{P}[\mathsf{X} = x | \mathsf{Z} = +1] \cdot \mathbb{P}[\mathsf{X}' = x' | \mathsf{Z} = +1] + \mathbb{P}[\mathsf{X} = x | \mathsf{Z} = -1] \cdot \mathbb{P}[\mathsf{X}' = x' | \mathsf{Z} = -1]} \\
&= \frac{1}{1 + \frac{\mathbb{P}[\mathsf{X} = x | \mathsf{Z} = -1] \cdot \mathbb{P}[\mathsf{X}' = x' | \mathsf{Z} = -1]}{\mathbb{P}[\mathsf{X} = x | \mathsf{Z} = +1] \cdot \mathbb{P}[\mathsf{X}' = x' | \mathsf{Z} = +1]}} \\
&= \sigma(\sigma^{-1}(\eta(x)) - \sigma^{-1}(\eta(x'))).
\end{aligned}$$

The last identity follows because

$$\sigma^{-1}(\eta(x)) = \log \frac{\pi}{1 - \pi} + \log \frac{\mathbb{P}[\mathsf{X} = x | \mathsf{Z} = +1]}{\mathbb{P}[\mathsf{X} = x' | \mathsf{Z} = -1]}.$$

■

From Lemma 37, the Bayes-optimal scorers for a classification calibrated loss are immediate.

Lemma 38 *For any $D = \langle M, \eta \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$, and any classification-calibrated ℓ,*

$$\mathcal{S}^*(D_{\text{BR}}, \ell) \subseteq \left\{ s_{\text{Pair}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R} : \eta(x) \neq \eta(x') \implies \text{sign}(s_{\text{Pair}}(x, x')) = \text{sign}(\eta(x) - \eta(x')) \right\}.$$

When ℓ is the 0-1 loss, the above is an equality.

Proof For a classification calibrated loss ℓ, for every $x, x' \in \mathcal{X}$ such that $\eta(x) \neq \eta(x')$, any Bayes-optimal pair-scorer $s^*_{\text{Pair}} \in \mathcal{S}^*(D_{\text{BR}}, \ell)$ must satisfy

$$\begin{aligned}\operatorname{sign}(s^*_{\text{Pair}}(x, x')) &= \operatorname{sign}(2\eta_{\text{Pair}}(x, x') - 1) \\ &= \operatorname{sign}(2\sigma(\sigma^{-1}(\eta(x)) - \sigma^{-1}(\eta(x'))) - 1) \text{ by Lemma 37} \\ &= \operatorname{sign}(\sigma^{-1}(\eta(x)) - \sigma^{-1}(\eta(x'))) \\ &= \operatorname{sign}(\eta(x) - \eta(x')). \end{aligned} \tag{48}$$

When $\eta(x) = \eta(x')$, we can pick any $s^*_{\text{Pair}}(x, x') \in \mathbb{R}$. Thus, for all $D = \langle M, \eta \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$ and classification-calibrated losses ℓ,

$$\mathcal{S}^*(D_{\text{BR}}, \ell) \subseteq \left\{ s_{\text{Pair}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \,:\, \eta(x) \neq \eta(x') \implies \operatorname{sign}(s_{\text{Pair}}(x, x')) = \operatorname{sign}(\eta(x) - \eta(x')) \right\}. \tag{49}$$

For ℓ_{01}, every pair-scorer satisfying the above is optimal, thus yielding an equality. ■

When ℓ is a proper composite loss with link Ψ, by definition we can specify one Bayes-optimal pair-scorer, though there may be others:

$$\{\Psi \circ \eta_{\text{Pair}}\} = \{\Psi \circ \sigma \circ \text{Diff}(\sigma^{-1} \circ \eta)\} \subseteq \mathcal{S}^*(D_{\text{BR}}, \ell). \tag{50}$$

This is an equality if and only if ℓ is strictly proper composite.

Having computed the optimal pair-scorers, when attempting to translate the results to bipartite ranking, we immediately face a challenge, as

$$\begin{aligned}\operatorname*{Argmin}_{s\,:\,\mathcal{X} \to \mathbb{R}} \mathbb{L}_{\text{BR}}(s; D, \ell) &= \operatorname*{Argmin}_{s\,:\,\mathcal{X} \to \mathbb{R}} \mathbb{L}(\text{Diff}(s); D_{\text{BR}}, \ell) \\ &= \operatorname*{Argmin}_{s_{\text{Pair}} \in \mathcal{S}_{\text{Decomp}}} \mathbb{L}(s_{\text{Pair}}; D_{\text{BR}}, \ell).\end{aligned}$$

That is, finding the set of scorers s that minimise $\mathbb{L}_{\text{BR}}(\text{Diff}(s))$ is equivalent to finding the set of pair-scorers s_{Pair} that minimise $\mathbb{L}(s_{\text{Pair}}; D_{\text{BR}}, \ell)$, *subject to* the pair-scorers being decomposable. Formally, in the notation of Equation 20, we need every such optimal $s^*_{\text{Pair}} \in \mathcal{L}_{\text{Decomp}}$. While the latter constraint seems innocuous, it means we need to reason about a minimiser in a *restricted* function class. Thus, in general, it is no longer possible to simply study the conditional risk and make a pointwise analysis.

Of course, we can easily make progress in the special case where the optimal pair-scorer is in fact decomposable – in this case, we can effectively ignore the restricted function class, because the optimal pair-scorer must be the difference of the optimal univariate scorer. The following makes this precise.

Proposition 39 *Given any $D \in \Delta_{\mathcal{X} \times \{\pm 1\}}$ and loss ℓ,*

$$\ell \in \mathcal{L}_{\text{Decomp}} \iff \mathcal{S}^*(D_{\text{BR}}, \ell) \cap \mathcal{S}_{\text{Decomp}} = \text{Diff}(\mathcal{S}^*_{\text{BR}}(D, \ell)).$$

Proof The ($\Longleftarrow$) direction is immediate, since $\mathcal{S}^*_{\text{BR}}(D, \ell) \neq \emptyset$ and thus $\text{Diff}(\mathcal{S}^*_{\text{BR}}(D, \ell)) \neq \emptyset$. We show the ($\Longrightarrow$) direction.

($\subseteq$). Pick any $s^*_{\text{Pair}} \in \mathcal{S}^*(D_{\text{BR}}, \ell) \cap \mathcal{S}_{\text{Decomp}}$. Then $s^*_{\text{Pair}} = \text{Diff}(s)$ for some $s : \mathcal{X} \to \mathbb{R}$. By optimality of s^*_{Pair},

$$(\forall t : \mathcal{X} \to \mathbb{R})\, \mathbb{L}_{\text{BR}}(s) = \mathbb{L}(s^*_{\text{Pair}}; D_{\text{BR}}, \ell) \leq \mathbb{L}(\text{Diff}(t); D_{\text{BR}}, \ell) = \mathbb{L}_{\text{BR}}(t).$$

Thus $s \in \mathcal{S}^*_{\text{BR}}(D, \ell)$, and so $s^*_{\text{Pair}} \in \text{Diff}(\mathcal{S}^*_{\text{BR}}(D, \ell))$.

($\supseteq$). Pick any $s^* \in \mathcal{S}^*_{\text{BR}}(D, \ell)$, and let $s_{\text{Pair}} = \text{Diff}(s^*)$. Then, by definition,

$$s_{\text{Pair}} \in \operatorname*{Argmin}_{t_{\text{Pair}} \in \mathcal{S}_{\text{Decomp}}} \mathbb{L}(t_{\text{Pair}}; D_{\text{BR}}, \ell).$$

This is a constrained optimisation problem. When $\ell \in \mathcal{L}_{\text{Decomp}}$, there is at least one solution to the *un*constrained optimisation that lies in $\mathcal{S}_{\text{Decomp}}$, call it t_{Pair}. Clearly t_{Pair} is a feasible solution for the constrained problem above. Thus, it must have an identical risk to s_{Pair}. But then s_{Pair} is a solution to the unconstrained problem as well, and so $s_{\text{Pair}} \in \mathcal{S}^*(D_{\text{BR}}, \ell) \cap \mathcal{S}_{\text{Decomp}}$. ■

The result simplifies somewhat when *every* Bayes-optimal pair-scorer is decomposable, which occurs when there is a unique optimal pair-scorer.

Corollary 40 *Given any $D \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ and loss ℓ,*

$$\mathcal{S}^*(D_{\text{BR}}, \ell) \subseteq \mathcal{S}_{\text{Decomp}} \iff \mathcal{S}^*(D_{\text{BR}}, \ell) = \text{Diff}(\mathcal{S}^*_{\text{BR}}(D, \ell)).$$

Proof ($\implies$) follows by Proposition 39, and ($\impliedby$) follows by definition of decomposability. ■

Simply put, the decomposable Bayes-optimal pair-scorers are exactly the Bayes-optimal scorers passed through $\text{Diff}(\cdot)$. Thus, if we can show that $\mathcal{S}^*_{\text{BR}}(D, \ell) \cap \mathcal{S}_{\text{Decomp}} \neq \emptyset$ for a loss ℓ, we automatically deduce the Bayes-optimal scorer from the results of the previous section. We determine when this condition holds below.

7.4 Bipartite ranking: decomposable case

We study the Bayes-optimal scorers for losses that induce a decomposable Bayes-optimal pair-scorer. We begin with the case ℓ_{01}.

7.4.1 Optimal univariate scorer for 0-1 loss

For ℓ_{01}, it is not hard to see that our earlier results imply that D_{BR} has at least one decomposable Bayes-optimal pair-scorer.

Lemma 41 *Given any $D = \langle M, \eta \rangle \in \Delta_{\mathcal{X}\times\{\pm 1\}}$,*

$$\mathcal{S}^*(D_{\text{BR}}, \ell_{01}) \cap \mathcal{S}_{\text{Decomp}} \neq \emptyset.$$

Proof By Equation 49, we see that $\{\text{Diff}(\eta)\} \subseteq \mathcal{S}^*(D_{\text{BR}}, \ell_{01}) \cap \mathcal{S}_{\text{Decomp}}$. That is, ℓ_{01} induces at least one decomposable Bayes-optimal pair-scorer in the pairwise ranking risk. ■

We can now show that the optimal scorers for the 0-1 bipartite risk are those that preserve the ordering of the class-probability η, which includes all strictly monotone transformations of η.

Proposition 42 *Given any $D = \langle M, \eta \rangle \in \Delta_{\mathcal{X}\times\{\pm 1\}}$,*

$$\mathcal{S}^*_{\text{BR}}(D, \ell_{01}) = \{s : \mathcal{X} \to \mathbb{R} : \eta = \phi \circ s \text{ for } \phi : \mathbb{R} \to [0,1] \text{ non-decreasing}\}.$$

Proof Let $\mathcal{A} = \mathcal{S}^*(D_{\text{BR}}, \ell_{01}) \cap \mathcal{S}_{\text{Decomp}}$. Since $\mathcal{A}$ is nonempty by Proposition 41, $\mathcal{A} = \text{Diff}(\mathcal{S}^*_{\text{BR}}(D, \ell_{01}))$ by Proposition 39. Equivalently, by Lemma 38,

$$\begin{aligned}
\mathcal{A} &= \left\{s_{\text{Pair}} \in \mathcal{S}_{\text{Decomp}} : \eta(x) \neq \eta(x') \implies \text{sign}(s_{\text{Pair}}(x, x')) = \text{sign}(\eta(x) - \eta(x'))\right\} \\
&= \text{Diff}\left(\left\{s : \mathcal{X} \to \mathbb{R} : \eta(x) \neq \eta(x') \implies \text{sign}(s(x) - s(x')) = \text{sign}(\eta(x) - \eta(x'))\right\}\right) \\
&= \text{Diff}\left(\{s : \mathcal{X} \to \mathbb{R} : \eta = \phi \circ s \text{ for non-decreasing } \phi\}\right) \text{ by Lemma 72.}
\end{aligned}$$

We thus have equality of the differences of the sets of interest. But for any sets of scorers $\mathcal{S}_1, \mathcal{S}_2$, $\text{Diff}(\mathcal{S}_1) = \text{Diff}(\mathcal{S}_2) \implies (\forall s_1 \in \mathcal{S}_1)(\exists s_2 \in \mathcal{S}_2, c \in \mathbb{R})\, s_1 = s_2 + c$, i.e. the scorers in the two sets must be related by a

linear translation. But if for a scorer s we have $\eta = \phi \circ s$ for some monotone ϕ, then it must also be true that $\eta = \tilde{\phi} \circ (s + c)$ where $\tilde{\phi} : x \mapsto \phi(x - c)$ is also monotone. Thus, the result follows. ■

The transform ϕ in Proposition 42 is not required to be strictly monotone increasing since if $\eta(x) = \eta(x')$ for some $x \neq x' \in \mathcal{X}$, it is allowed for $s(x) \neq s(x')$. (In the extreme case where $\eta(x) \equiv c$ for every x, then every scorer will trivially be Bayes-optimal.) Nonetheless, an immediate corollary is that any strictly monotone increasing transform of η is necessarily an optimal univariate scorer[15].

Corollary 43 *Given any $D = \langle M, \eta \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$ and any strictly monotone increasing $\phi : [0, 1] \to \mathbb{R}$,*

$$\phi \circ \eta \in \mathcal{S}^*_{\mathrm{BR}}(D, \ell_{01}).$$

We see that like class-probability estimation, bipartite ranking with ℓ_{01} aims to find a transformation of η. Unlike class-probability estimation, one is satisfied with *any* strictly monotone transformation, not necessarily one specified by the loss itself. Loosely, then, bipartite ranking is less "strict" than class-probability estimation. (See also §10.)

7.4.2 Optimal univariate scorer for strictly proper composite losses

We now proceed to the case where ℓ is a strictly proper composite loss. To apply Corollary 40, we characterise the subset of proper composite losses for which there exists a decomposable pair-scorer. This shall turn out to rely on the following set of inverse link functions from the sigmoid family,

$$\Sigma_{\mathrm{sig}} \doteq \left\{ \Psi^{-1} : \mathbb{R} \to [0, 1] \mid (\exists a \in \mathbb{R} \setminus \{0\}) (\forall v \in \mathbb{R}) \, \Psi^{-1}(v) = \frac{1}{1 + e^{-av}} \right\}. \tag{51}$$

Proposition 44 (Decomposability of Bayes-optimal bipartite pair-scorer.) *Given any $\ell \in \mathcal{L}_{\mathrm{SPC}}(\Psi)$ with Ψ differentiable,*

$$(\forall D \in \Delta_{\mathcal{X} \times \{\pm 1\}}) \, \mathcal{S}^*(D_{\mathrm{BR}}, \ell) \subseteq \mathcal{S}_{\mathrm{Decomp}} \iff \Psi^{-1} \in \Sigma_{\mathrm{sig}}.$$

Proof ($\Longleftarrow$) Let the link function of ℓ have the specified form, so that $\Psi(v) = \frac{1}{a} \log \frac{v}{1-v} = \frac{1}{a} \sigma^{-1}(v)$, and so $(\Psi \circ \sigma)(v) = \frac{v}{a}$. From Equation 50, the[16] Bayes-optimal pair-scorer is

$$\begin{aligned} s^*_{\mathrm{Pair}} &= \frac{1}{a} \cdot \mathrm{Diff}(\sigma^{-1} \circ \eta) \\ &= \mathrm{Diff}\left(\left(\frac{1}{a} \cdot \sigma^{-1} \right) \circ \eta \right) \\ &\in \mathcal{S}_{\mathrm{Decomp}}. \end{aligned}$$

Thus $s^*_{\mathrm{Pair}} \in \mathcal{S}^*(D_{\mathrm{BR}}, \ell) \cap \mathcal{S}_{\mathrm{Decomp}}$.

($\Longrightarrow$) The proof here uses a similar idea to Uematsu and Lee (2012, Theorem 7). If $\ell \in \mathcal{L}_{\mathrm{Decomp}}$,

$$\Psi \circ \sigma \circ \mathrm{Diff}(\sigma^{-1} \circ \eta) \in \mathcal{S}_{\mathrm{Decomp}}.$$

We wish to determine the nature of Ψ that permits this to hold. Let $f = \Psi \circ \sigma \circ \log$, so that the above becomes

$$(\forall x, x' \in \mathcal{X}) \, f\left(\frac{e^{\sigma^{-1}(\eta(x))}}{e^{\sigma^{-1}(\eta(x'))}} \right) = g(x) - g(x')$$

15. Combined with the connection between the 0-1 optimal scorer for the bipartite ranking risk and AUC (Corollary 23), this constitutes an alternate proof of Corollary 16, without an appeal to the Neyman-Pearson lemma.

16. For a *non-strict* proper composite loss, the following argument holds, but only for one possible optimal pair-scorer. Thus, it may not be true that *all* Bayes-optimal pair-scorers are decomposable; however, for the choice of link function above, we can guarantee that there is *at least* one that is. Nonetheless, for a *strictly* proper composite loss, there is a unique Bayes-optimal pair-scorer. Thus, the above result characterises when this pair-scorer is decomposable.

for some $g : \mathcal{X} \to \mathbb{R}$. Now note that

$$(\forall x, x', x'' \in \mathcal{X})\, f\left(\frac{e^{\sigma^{-1}(\eta(x))}}{e^{\sigma^{-1}(\eta(x'))}}\right) = g(x) - g(x'') + g(x'') - g(x') = f\left(\frac{e^{\sigma^{-1}(\eta(x))}}{e^{\sigma^{-1}(\eta(x''))}}\right) + f\left(\frac{e^{\sigma^{-1}(\eta(x''))}}{e^{\sigma^{-1}(\eta(x'))}}\right).$$

We require this to hold for any D, and thus for any η. Therefore, equivalently, we have

$$(\forall a, b \in \mathbb{R}_+)\, f(a \cdot b) = f(a) + f(b).$$

Note that f is continuous by assumed differentiability of Ψ. Thus the only solution to the equation is $f(z) = \frac{1}{a} \cdot \log z$ for some $a \in \mathbb{R}$ (Kannappan, 2009, Corollary 1.43), or equivalently that $\Psi^{-1}(v) = \sigma(a \cdot v) = \frac{1}{1+e^{-av}}$. Note that the case $a = 0$ is ruled out by assumed invertibility of Ψ, and thus equivalently of f. ■

We emphasise that the class of proper composite losses satisfying the above condition is "large" in the following sense: one may take *any* strictly proper loss and compose it with any member of the given link family. Two specific implications are noteworthy. First, the loss ℓ need not be symmetric; Appendix G has an empirical illustration of this fact. Second, the loss ℓ may be non-convex; nonetheless, we can easily determine the optimal scorers for all such losses, as below.

Corollary 45 *Given any $D = \langle M, \eta \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$ and $\ell \in \mathcal{L}_{\text{SPC}}(\Psi)$ with inverse link function $\Psi^{-1} \in \Sigma_{\text{sig}}$,*

$$\mathcal{S}^*_{\text{BR}}(D, \ell) = \{\Psi \circ \eta + b : b \in \mathbb{R}\}.$$

Consequently, when Ψ is monotone increasing (viz. when $a \in \mathbb{R}_+$ in Equation 51),

$$\mathcal{S}^*_{\text{BR}}(D, \ell) \subseteq \mathcal{S}^*_{\text{BR}}(D, \ell_{01}).$$

Proof By Proposition 44 and Corollary 40,

$$\text{Diff}(\mathcal{S}^*_{\text{BR}}(D, \ell)) = \mathcal{S}^*_{\text{BR}}(D, \ell).$$

Further, by Equation 50, and letting $\Psi^{-1} : v \mapsto \sigma(a \cdot v)$,

$$\mathcal{S}^*_{\text{BR}}(D, \ell) = \text{Diff}\left(\frac{1}{a} \cdot \sigma^{-1} \circ \eta\right) = \text{Diff}(\Psi \circ \eta).$$

The result follows because

$$\text{Diff}(f) = \text{Diff}(g) \iff (\exists b \in \mathbb{R})\, f = g + b.$$

■

The admissible family of links Σ_{sig} can be easily checked to contain those employed for the logistic and exponential losses, and thus we can deduce the decomposability of the Bayes-optimal scorers for these losses.

Corollary 46 *For the strictly proper composite logistic and exponential losses,*

$$\ell_{\log}(y, z) = \log\left(1 + e^{-yz}\right)$$
$$\ell_{\exp}(y, z) = e^{-yz},$$

with inverse link functions $\Psi^{-1}(v) = \frac{1}{1+e^{-v}}$ and $\Psi^{-1}(v) = \frac{1}{1+e^{-2v}}$ respectively, the Bayes-optimal pair-scorer is decomposable for any $D \in \Delta_{\mathcal{X} \times \{\pm 1\}}$ i.e.

$$\mathcal{S}^*(D_{\text{BR}}, \ell) \subseteq \mathcal{S}_{\text{Decomp}}.$$

While Proposition 44 follows easily from the proper loss machinery, the requirement on the link function is *a priori* non-obvious. What is special about link functions that are scaled versions of the sigmoid? The answer is simply that the scorer η_{Pair} inherently involves a sigmoid link function (Lemma 37). This form of η_{Pair} in turn can be understood via utility representations for binary relations on sets, as we discuss in §12.

7.4.3 COMMENT ON CONVEXITY

In general, an invertible link function Ψ can be composed with any proper loss to yield a proper composite loss. For numerical convenience, it is useful to consider only those proper losses which yield a *convex* proper composite loss. For a proper loss λ, let $\ell(y, v) = \lambda(y, \Psi^{-1}(v))$ for $\Psi^{-1}(v) \in \Sigma_{\text{sig}}$ i.e. $\Psi^{-1}(v) = \frac{1}{1+e^{-av}}$ for some a. Such a loss will have Bayes-optimal scorer as given by Proposition 44; but when will such a loss be additionally convex? Suppose the weight function w for λ is normalised such that $w\left(\frac{1}{2}\right) = 1$. Then, ℓ is convex only if the weight function w satisfies (Reid and Williamson, 2010, Theorem 29)

$$w(c) \in \left[\min\left(\frac{1}{a \cdot c^2 \cdot (1-c)}, \frac{1}{a \cdot c \cdot (1-c)^2}\right), \max\left(\frac{1}{a \cdot c^2 \cdot (1-c)}, \frac{1}{a \cdot c \cdot (1-c)^2}\right)\right].$$

The above gives necessary conditions[17] for obtaining a convex proper composite loss with the given link.

As a sanity check, two losses encountered earlier in Corollary 46 will indeed satisfy the above. For the admissible weight function $w(c) = \frac{1}{a \cdot (c \cdot (1-c))^{3/2}}$, it is easy to check that with a sigmoidal link we recover a generalised version of the exponential loss, $\ell(y, v) = \frac{1}{a} e^{-yav}$. (We will revisit these family of losses in a different context in §9.5.) Recall from Equation 13 that a link Ψ is canonical for a given proper loss λ with weight function w when $w(c) = \Psi'(c)$. For $\Psi^{-1}(v) = \frac{1}{1+e^{-av}}$, we have $w(c) = \frac{1}{a} \cdot \left(\log \frac{c}{1-c}\right)' = \frac{1}{a \cdot c \cdot (1-c)}$. The resulting proper composite loss is

$$\ell(y, v) = \lambda(y, \Psi^{-1}(v)) = \frac{1}{a} \cdot \log(1 + e^{-yav}),$$

which is a generalised logistic loss. (Masnadi-Shirazi and Vasconcelos (2010) call this the canonical logistic loss.) Note that $\lim_{a \to \infty} \frac{1}{a} \log(1 + e^{-yav}) = \max(0, -v)$, which is the perceptron loss.

7.5 Bipartite ranking: non-decomposable case

We now turn to the case where the loss ℓ does *not* have a decomposable Bayes-optimal pair-scorer. As noted earlier, we can no longer resort to reasoning solely via the conditional risk. Fortunately, the simple structure of $\mathcal{S}_{\text{Decomp}}$ means that we can hope to directly compute the risk minimiser via an appropriate derivative. Under some assumptions on the loss, it turns out that the Bayes-optimal scorer is still a strictly monotone transform of η; however, the transform is now *distribution dependent*, rather than simply the fixed link function Ψ.

Proposition 47 *Pick any $D = \langle M, \eta \rangle = \langle P, Q, \pi \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$ and a differentiable, convex, symmetric strictly proper composite loss $\ell(y, v) = \phi(yv)$. If ϕ' is bounded[18], or the support of D is finite,*

$$\mathcal{S}^*_{\text{BR}}(D, \ell) = \{s^* : \mathcal{X} \to \mathbb{R} \mid \eta = f_{D,s^*} \circ s^*\},$$

where

$$(\forall v \in \mathcal{V})\, f_{D,s^*}(v) \doteq \frac{\pi \cdot \mathop{\mathbb{E}}_{\mathsf{X} \sim P}\left[\ell'_{-1}(v - s^*(\mathsf{X}))\right]}{\pi \cdot \mathop{\mathbb{E}}_{\mathsf{X} \sim P}\left[\ell'_{-1}(v - s^*(\mathsf{X}))\right] - (1-\pi) \cdot \mathop{\mathbb{E}}_{\mathsf{X}' \sim Q}\left[\ell'_{1}(v - s^*(\mathsf{X}'))\right]}.$$

17. More complex sufficient conditions may also be derived; see (Reid and Williamson, 2010, Theorem 24).
18. We suspect this requirement may be dropped, but defer to future work investigation of minimal conditions for the result to hold.

Proof The basic proof strategy follows Uematsu and Lee (2012, Theorem 3), although the subsequent steps and connection to proper loss concepts are novel; we will shortly discuss the connection to the results of that paper.

Let $\ell(y, v) = \phi(yv)$. For fixed D, let $\mathcal{S}(D)$ denote the space of all Lebesgue-measurable scorers $s : \mathcal{X} \to \mathbb{R}$, with addition and scalar multiplication defined pointwise, such that

$$\mathbb{L}_{\text{BR}}(s; D, \ell) = \mathop{\mathbb{E}}_{\mathsf{X} \sim P, \mathsf{X}' \sim Q} \left[\phi(s(\mathsf{X}) - s(\mathsf{X}'))\right] < \infty.$$

Then $\mathbb{L}_{\text{BR}} : \mathcal{S}(D) \to \mathbb{R}$ is a convex functional, by virtue of ϕ being convex. Thus, its minimisers may be determined by considering an appropriate notion of functional derivative. We shall employ the Gâteaux variation[19].

Pick any $s, t \in \mathcal{S}(D)$. For any $\epsilon > 0$, define

$$\begin{aligned} F_{s,t}(\epsilon) &= \mathbb{L}_{\text{BR}}(s + \epsilon t) \\ &= \mathop{\mathbb{E}}_{\mathsf{X} \sim P, \mathsf{X}' \sim Q} \left[\phi(s(\mathsf{X}) - s(\mathsf{X}') + \epsilon(t(\mathsf{X}) - t(\mathsf{X}')))\right]. \end{aligned}$$

The Gâteaux variation of $\mathbb{L}_{\text{BR}}$ at s in the direction of t is (Troutman, 1996, pg. 45), (Giaquinta and Hildebrandt, 2004, pg. 10)

$$\begin{aligned} \delta\mathbb{L}_{\text{BR}}(s; t) &\doteq \lim_{\epsilon \to 0} \frac{\mathbb{L}_{\text{BR}}(s + \epsilon t) - \mathbb{L}_{\text{BR}}(s; D, \ell)}{\epsilon} \\ &= F'_{s,t}(0), \end{aligned}$$

assuming the latter exists. To show that $F'_{s,t}(0)$ exists, we will justify interchange of the derivative and expectation. For any $\epsilon \in (0, 1]$ and $x, x' \in \mathcal{X}$, by convexity and nonnegativity of ϕ,

$$\begin{aligned} \left|\frac{\phi((\text{Diff}(s + \epsilon t))(x, x')) - \phi((\text{Diff}(s))(x, x'))}{\epsilon}\right| &\leq \left|\phi((\text{Diff}(s + t))(x, x')) - \phi((\text{Diff}(s))(x, x'))\right| \\ &\leq \phi((\text{Diff}(s + t))(x, x')) + \phi((\text{Diff}(s))(x, x')). \end{aligned}$$

By assumption, $\mathbb{L}_{\text{BR}}(s + t)$ and $\mathbb{L}_{\text{BR}}(s; D, \ell)$ are both finite. Further,

$$\lim_{\epsilon \to 0} \frac{\phi(s(x) - s(x') + \epsilon(t(x) - t(x'))) - \phi(s(x) - s(x'))}{\epsilon} = (t(x) - t(x')) \cdot \phi'(s(x) - s(x')).$$

Thus, by the dominated convergence theorem (Folland, 1999, pg. 56), we have

$$\begin{aligned} F'_{s,t}(0) &= \mathop{\mathbb{E}}_{\mathsf{X} \sim P, \mathsf{X}' \sim Q} \left[(t(\mathsf{X}) - t(\mathsf{X}')) \cdot \phi'(s(\mathsf{X}) - s(\mathsf{X}'))\right] \\ &= \mathop{\mathbb{E}}_{\mathsf{X} \sim P, \mathsf{X}' \sim Q} \left[t(\mathsf{X}) \cdot \phi'(s(\mathsf{X}) - s(\mathsf{X}'))\right] - \mathop{\mathbb{E}}_{\mathsf{X} \sim Q, \mathsf{X}' \sim P} \left[t(\mathsf{X}) \cdot \phi'(s(\mathsf{X}') - s(\mathsf{X}))\right] \\ &= \int_{\mathcal{X}} t(x) \cdot r(x) \, dx, \end{aligned}$$

where

$$(\forall x \in \mathcal{X})\, r(x) \doteq p(x) \cdot \mathbb{E}_{\mathsf{X}' \sim Q} \left[\phi'(s(x) - s(\mathsf{X}'))\right] - q(x) \cdot \mathbb{E}_{\mathsf{X} \sim P} \left[\phi'(s(\mathsf{X}) - s(x))\right].$$

Now suppose $s^* : \mathcal{X} \to \mathbb{R}$ minimises the functional $\mathbb{L}_{\text{BR}}$. By convexity of $\mathbb{L}_{\text{BR}}$, it is necessary and sufficient that the Gâteaux variation is zero for $t \in \mathcal{L}(D)$ (Gelfand and Fomin, 2000, Theorem 2), (Troutman, 1996, Proposition 3.3). That is,

$$(\forall t \in \mathcal{L}(D))\, 0 = \int_{\mathcal{X}} t(x) \cdot r(x) \, dx.$$

19. This is distinct from the Gâteaux *derivative*, which in turn is distinct from the Fréchet derivative. The latter concepts require that we define a norm over the input space of the functional. A norm is needed to define a local extrema of a functional. As our $\mathbb{L}_{\text{BR}}$ is convex, the extrema we determine will be local extrema according to any norm (Troutman, 1996, pg. 114).

A sufficient condition for this to hold is that r is zero (almost) everywhere, and this is in fact necessary as well (Lemma 71). That is, for (almost) every $x_0 \in \mathcal{X}$, we equivalently need

$$p(x_0) \cdot \mathbb{E}_{X' \sim Q}\left[\phi'(s^*(x_0) - s^*(X'))\right] = q(x_0) \cdot \mathbb{E}_{X \sim P}\left[\phi'(s^*(X) - s^*(x_0))\right],$$

which means for (almost) every $x_0 \in \mathcal{X}$,

$$\begin{aligned}
\frac{\eta(x_0)}{1-\eta(x_0)} \cdot \frac{1-\pi}{\pi} &= \frac{p(x_0)}{q(x_0)} \\
&= \frac{\mathbb{E}_{X \sim P}\left[\phi'(s^*(X) - s^*(x_0))\right]}{\mathbb{E}_{X' \sim Q}\left[\phi'(s^*(x_0) - s^*(X'))\right]} \\
&= \frac{\mathbb{E}_{X \sim P}\left[\ell'_1(s^*(X) - s^*(x_0)) - \ell'_{-1}(s^*(x_0) - s^*(X))\right]}{\mathbb{E}_{X' \sim Q}\left[-\ell'_1(s^*(x_0) - s^*(X')) + \ell'_{-1}(s^*(X) - s^*(x_0))\right]} \\
&= \frac{\mathbb{E}_{X \sim P}\left[\ell'_{-1}(s^*(x_0) - s^*(X)) - \ell'_1(s^*(X) - s^*(x_0))\right]}{\mathbb{E}_{X' \sim Q}\left[\ell'_1(s^*(x_0) - s^*(X')) - \ell'_{-1}(s^*(X) - s^*(x_0))\right]} \\
&= \frac{\mathbb{E}_{X \sim P}\left[\ell'_{-1}(s^*(x_0) - s^*(X))\right]}{\mathbb{E}_{X' \sim Q}\left[\ell'_1(s^*(x_0) - s^*(X'))\right]} \text{ since } \ell \text{ is symmetric,}
\end{aligned}$$

which means

$$\eta = f_{D,s^*} \circ s^*,$$

where f_{D,s^*} is given by

$$(f_{D,s^*})(v) \doteq \frac{\pi \cdot \mathbb{E}_{X \sim P}\left[\ell'_{-1}(v - s^*(X))\right]}{\pi \cdot \mathbb{E}_{X \sim P}\left[\ell'_{-1}(v - s^*(X))\right] - (1-\pi) \cdot \mathbb{E}_{X' \sim Q}\left[\ell'_1(v - s^*(X'))\right]}.$$

■

In order to express any optimal scorer s^* in terms of η, as we have done for the previous cases, it remains to check whether or not the above the function f_{D,s^*} defined above is invertible. The following corollary provides sufficient conditions for this to hold.

Corollary 48 *Pick any $D = \langle M, \eta \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$ and margin-based strictly proper composite loss $\ell(y, v) = \phi(yv)$, where ϕ is differentiable, strictly convex, and satisfies*

$$(\forall v \in \mathbb{R})\, \phi'(v) = 0 \iff \phi'(-v) \neq 0.$$

Then if ϕ' is bounded or the support of D is finite, f_{D,s^*} is invertible and

$$\mathcal{S}^*_{\mathrm{BR}}(D, \ell) = \{s^* : \mathcal{X} \to \mathbb{R} \mid s^* = (f_{D,s^*})^{-1} \circ \eta\} \subseteq \mathcal{S}^*_{\mathrm{BR}}(D, \ell_{01}),$$

where f_{D,s^*} is defined as in Proposition 47.

Proof We show that f_{D,s^*} strictly monotone, by establishing the strict monotonicity of

$$g : v \mapsto \frac{\mathbb{E}_{X' \sim Q}\left[\ell'_1(v - s^*(X'))\right]}{\mathbb{E}_{X \sim P}\left[\ell'_{-1}(v - s^*(X))\right]}.$$

The derivative of this function is

$$g'(v) = \frac{1}{\left(\mathop{\mathbb{E}}_{\mathsf{X}\sim P}\left[\ell'_{-1}(v - s^*(\mathsf{X}))\right]\right)^2} \cdot \left(\mathop{\mathbb{E}}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[\ell'_{-1}(v - s^*(\mathsf{X}))\ell''_1(v - s^*(\mathsf{X}')) - \ell''_{-1}(v - s^*(\mathsf{X}))\ell'_1(v - s^*(\mathsf{X}'))\right]\right).$$

By convexity of ℓ, the terms $\ell''_1(v - s^*(\mathsf{X}'))$ and $\ell''_{-1}(v - s^*(\mathsf{X}))$ are positive. Further, by (Vernet et al., 2011, Proposition 15), ℓ_1 and ℓ_{-1} are increasing and decreasing respectively, or vice-versa. By assumption their derivatives cannot simultaneously be zero. Therefore the expectand is always positive or negative for every v, and hence $g'(v)$ is always strictly positive or negative. Thus g is strictly monotone, which means f_{D,s^*} is as well. Therefore, $s^* = (f_{D,s^*})^{-1} \circ \eta$. ■

The link function f_{D,s^*} is at first glance peculiar because it depends on the distribution D, as well as the optimal scorer s^*. From a practical perspective, the result is thus not helpful in terms of helping quickly discover the optimal scorer s^*. However, what is of interest to us is simply that this function is strictly monotone. This means that from an AUC perspective, using a proper composite surrogate loss asymptotically results in a desirable scorer i.e. if we rank examples according to s^*, it is equivalent to ranking them according to η.

As before, any optimal scorer for such a proper composite loss is also optimal for ℓ_{01}, despite the link function f_{D,s^*} depending on the distribution D. Appendix H provides an empirical illustration that this link is indeed invertible under the specified conditions, albeit distribution dependent. The results of this section established that a suitably restricted notion of convexity is *sufficient* for the optimal scorer to be a strictly monotone transform of η, while the previous section established convexity is *not necessary*, since one can have a non-convex loss resulting from a suitable link $\Psi = \frac{1}{a}\sigma^{-1}$.

7.5.1 COMPARISON WITH STANDARD LINK FUNCTION

Recall from Equation 12 that the link function Ψ associated with a proper composite loss ℓ satisfies

$$\Psi^{-1}(v) = \frac{\ell'_{-1}(v)}{\ell'_{-1}(v) - \ell'_1(v)}.$$

This is in general *not* the same as the inverse link function $(f_{D,s^*})^{-1}$ from the above result for a simple reason: the latter potentially depends on the distribution D, and the optimal scoring function s^* itself. However, the forms of the two functions are closely related: in $(f_{D,s^*})^{-1}$, each quantity from the inverse link Ψ^{-1} is replaced by its expected value under an appropriate distribution.

In the previous section, we saw that under certain conditions $s^* = \Psi \circ \eta$, where Ψ is the standard link function associated with ℓ. In such cases, it is of interest to see whether f_{D,s^*} simplifies. For example, for the case of exponential loss $\ell(y, v) = e^{-yv}$, with inverse link function $\Psi^{-1}(v) = \frac{1}{1+e^{-2v}}$, we get

$$(f_{D,s^*})^{-1}(v) = \frac{\mathop{\mathbb{E}}_{\mathsf{X}\sim M}\left[\eta(\mathsf{X}) \cdot e^{s^*(\mathsf{X})}\right]}{\mathop{\mathbb{E}}_{\mathsf{X}\sim M}\left[\eta(\mathsf{X}) \cdot e^{s^*(\mathsf{X})}\right] + e^{-2v} \cdot \mathop{\mathbb{E}}_{\mathsf{X}\sim M}\left[(1 - \eta(\mathsf{X})) \cdot e^{-s^*(\mathsf{X})}\right]}.$$

It can be verified that when plugging in $s^* = \Psi \circ \eta$, one finds $(f_{D,s^*})^{-1}(v) = \Psi^{-1}(v)$ as expected, and therefore the dependence on the distribution "disappears". Determining conditions beyond $\Psi \in \Sigma_{\text{sig}}$ for which f_{D,s^*} simplifies would be of interest.

In class-probability estimation with a proper composite loss, there is a separation of concerns between the underlying proper loss and the link function Ψ, with the latter primarily chosen for computational convenience, and not affecting statistical properties of the proper loss (Reid and Williamson, 2010). For bipartite ranking,

however, such a separation of concerns is guaranteed only when one operates with the family of link functions from Proposition 44. For this family, the Bayes-optimal scorer is any translation of $\Psi \circ \eta$, while Corollary 48 indicates that outside this family, the optimal scorer may be a distribution-dependent transformation of η. Thus, changing the link function in bipartite ranking can change the optimal solutions to the risk in a non-trivial way.

7.6 Relation to existing work

The study of Bayes-optimal scorers for pairwise and bipartite ranking problems does not seem as extensive as for the binary classification setting. The study of the Bayes-optimal scorers for both problems under proper composite losses appears to be novel, although Agarwal (2014) employs proper losses in theoretical analysis related to the bipartite ranking problem. Even our derivation of the form of η_{Pair} does not have many precedents, though Uematsu and Lee (2012) implicitly derive the formula.

This section generalised and unified several earlier results through the theory of proper losses. For ℓ_{01}, our Corollary 43 is well-known in the context of scorers that maximise the AUC, which is one minus the bipartite ℓ_{01} risk. The result is typically established by the Neyman-Pearson lemma (Torgersen, 1991), whereas we simply use a reduction to binary classification over pairs. For exponential loss with a linear hypothesis class, Ertekin and Rudin (2011) studied the (empirical) Bayes-optimal solutions. For a convex margin loss, Uematsu and Lee (2012) and Gao and Zhou (2012, 2015) independently studied conditions for the Bayes-optimal scorers to be transformations of η. Our Proposition 44 is a generalisation of Uematsu and Lee (2012, Theorem 7), Gao and Zhou (2012, Lemma 3) and Gao and Zhou (2015, Corollary 1), where our result holds for non-symmetric and non-convex proper composite losses; Appendix G has an empirical illustration of this. Our Corollary 48 is essentially equivalent to Uematsu and Lee (2012, Theorem 3), Gao and Zhou (2012, Theorem 5), and Gao and Zhou (2015, Theorem 2), although we explicitly provide the form of the link function relating η and s^*. (We translate these results in terms of proper losses so that the connection is more apparent in Appendix F.)

Uematsu and Lee (2012, Theorem 5) also showed that for the hinge loss, which is classification calibrated, there are possibly ties introduced in the ranking, which is not covered by our results as the hinge loss is not proper. Gao and Zhou (2012, Theorem 2, 3) and Gao and Zhou (2015, Lemma 3) similarly show that hinge and absolute loss do not produce a consistent ranking.

8. Surrogate regret for pairwise surrogate minimisation

At this stage, we have established that for suitable ℓ, minimisers of the ℓ-bipartite risk will also minimise the ℓ_{01} bipartite risk. Equivalently, these scorers will maximise the AUC. This can be seen as a justification for the minimisation of a pairwise surrogate to ℓ_{01} for the task of maximising the AUC; this is sometimes referred to as the pairwise approach to bipartite ranking.

In practice, of course, we cannot expect to perfectly minimise $\mathbb{L}_{\text{BR}}$, due to having access to a finite sample, and (or) using a restricted function class of scorers. Thus, we would ideally like a bound as to how much worse the AUC can be when using a *suboptimal* minimiser of $\mathbb{L}_{\text{BR}}(s; D, \ell)$. This is the analogue of *surrogate regret bounds* for classification, which establish that convex risk minimisation is consistent for the problem of 0-1 minimisation (Zhang, 2004; Bartlett et al., 2006; Reid and Williamson, 2009).

Surrogate regret bounds have previously been shown for bipartite ranking by Clémençon et al. (2008, Section 7), under the implicit assumption of decomposable scorers, and by Gao and Zhou (2012, Corollary 6, 7), Gao and Zhou (2015, Corollary 5) for the symmetric exponential and logistic losses. It turns out that the analysis of the previous section automatically implies the existence of surrogate regret bounds for pairwise minimisation of suitable (possibly asymmetric) proper composite losses. Formally, we have the following.

Proposition 49 *Pick any $D \in \Delta_{\mathcal{X} \times \{\pm 1\}}$ and scorer $s : \mathcal{X} \to \mathbb{R}$. Given any $\ell \in \mathcal{L}_{\text{SPC}}(\Psi)$ with inverse link function $\Psi^{-1} \in \Sigma^+_{\text{sig}}$, there exists a convex function $F_\ell : [0, 1] \to \mathbb{R}_+$ with $F(0) = 0$ such that,*

$$F_\ell \left(\text{regret}_{\text{BR}}(s; D, \ell_{01}) \right) \leq \text{regret}_{\text{BR}}(s; D, \ell).$$

Proof For any $s : \mathcal{X} \to \mathbb{R}$, $s_{\text{Pair}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, recall that

$$\begin{aligned}\text{regret}(s_{\text{Pair}}; D_{\text{BR}}, \ell) &= \mathbb{L}(s_{\text{Pair}}; D_{\text{BR}}, \ell) - \inf_{t_{\text{Pair}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}} \mathbb{L}(t_{\text{Pair}}; D_{\text{BR}}, \ell) \\ \text{regret}_{\text{BR}}(s; D, \ell) &= \mathbb{L}_{\text{BR}}(s; D, \ell) - \inf_{t : \mathcal{X} \to \mathbb{R}} \mathbb{L}_{\text{BR}}(t).\end{aligned}$$

Existing surrogate regret bounds for proper composite losses (Reid and Williamson, 2009) imply that there exists some convex $F_\ell : [0, 1] \to \mathbb{R}_+$ such that, for any $D \in \Delta_{\mathcal{X} \times \{\pm 1\}}$ and scorer $s : \mathcal{X} \to \mathbb{R}$,

$$F_\ell \left(\text{regret}(\text{Diff}(s); D_{\text{BR}}, \ell_{01}) \right) \leq \text{regret}(\text{Diff}(s); D_{\text{BR}}, \ell).$$

By the reduction of bipartite ranking to classification over pairs (Lemma 2), for any ℓ satisfying the conditions of the proposition,

$$\begin{aligned}\text{regret}_{\text{BR}}(s; D, \ell) &= \mathbb{L}_{\text{BR}}(s; D, \ell) - \inf_{t : \mathcal{X} \to \mathbb{R}} \mathbb{L}_{\text{BR}}(t) \\ &= \mathbb{L}(\text{Diff}(s); D_{\text{BR}}, \ell) - \inf_{t_{\text{Pair}} \in \mathcal{S}_{\text{Decomp}}} \mathbb{L}(t_{\text{Pair}}; D_{\text{BR}}, \ell) \\ &= \mathbb{L}(\text{Diff}(s); D_{\text{BR}}, \ell) - \inf_{t_{\text{Pair}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}} \mathbb{L}(t_{\text{Pair}}; D_{\text{BR}}, \ell) \\ &= \text{regret}(\text{Diff}(s); D_{\text{BR}}, \ell),\end{aligned}$$

where in the penultimate line we have used the fact that the restriction to $\mathcal{S}_{\text{Decomp}}$ can be removed by virtue of the loss ℓ inducing a decomposable Bayes-optimal pair-scorer for D_{BR} (by Proposition 44). We similarly know that ℓ_{01} induces a decomposable pair-scorer (Proposition 41). Thus, we can write the regret bound as

$$F_\ell \left(\text{regret}_{\text{BR}}(s; D, \ell_{01}) \right) \leq \text{regret}_{\text{BR}}(s; D, \ell).$$

■

The function $F_\ell : [0, 1] \to \mathbb{R}_+$ in Proposition 49 is exactly that which appears in bounds relating 0-1 to ℓ classification regret for proper composite losses, and may be specified in terms of the conditional Bayes-risk of ℓ as (Reid and Williamson, 2009, Theorem 3)

$$\begin{aligned}F_\ell &: u \mapsto L^* \left(\frac{1}{2} \right) + G_\ell(u) \vee G_\ell(-u) \\ G_\ell &: u \mapsto -L^* \left(\frac{1}{2} + u \right) + (L^*)' \left(\frac{1}{2} \right) \cdot u.\end{aligned}$$

We make three observations about this result. First, the bound implies the consistency of pairwise surrogate minimisation for losses satisfying the conditions of the proposition, and whose underlying proper loss λ additionally satisfy the regularity condition $L^*(0) = 0$, as

$$\begin{aligned}\text{regret}_{\text{BR}}(s; D, \ell) \to 0 &\implies F_\ell(\text{regret}_{\text{BR}}(s; D, \ell_{01})) \to 0 \\ &\implies \text{regret}_{\text{BR}}(s; D, \ell_{01}) \to 0,\end{aligned}$$

where the second line is because $L^*(u) > 0$ on $(0, 1/2]$ by strict concavity of the conditional Bayes risk (a consequence of strict properness of the loss), and $L^*(0) = 0$ by assumption.

Second, the bound places no convexity restriction on ℓ. This is akin to similar regret bounds for classification (Bartlett et al., 2006, Theorem 1), where the surrogate loss need not be convex. Of course, for non-convex ℓ, guaranteeing $\text{regret}_{\text{BR}}(s; D, \ell) \to 0$ is more challenging.

Third, when the optimal pair-scorer is *not* decomposable, the proof breaks when attempting to equate $\text{regret}_{\text{BR}}(s; D, \ell)$ and $\text{regret}(\text{Diff}(s); D_{\text{BR}}, \ell)$, and so more effort is needed to derive a surrogate regret bound. This further illustrates the value of the decomposability of the Bayes-optimal pair-scorer as studied in the previous section. Note that while we do not have a regret bound for such losses, Corollary 48 established a

sufficient condition for agreement of the Bayes-optimal scorers. Further, Gao and Zhou (2012, Theorem 2) showed that for a subset of such losses, one has asymptotic consistency of the surrogate minimisation (even when a regret bound is elusive).

As a final note, the above is distinct from Agarwal (2014) as the latter bounds the AUC regret in terms of the regret with respect to a proper composite loss. That is, the result shows the consistency of the class-probability estimation approach to bipartite ranking. This is a distinct to our bound, which shows the consistency of the surrogate pairwise ranking approach to bipartite ranking.

9. Ranking the best instances

In most practical applications of ranking, accuracy at the head of the ranked list is more important than accuracy at the tail. For example, in information retrieval, typically only the first few elements of the ranked result set for a query are considered by a user of the system. It is thus of interest to consider notions of risk that focus on accuracy at the head of the list. This problem is sometimes called *ranking the best* (Clémençon and Vayatis, 2007) or *accuracy at the top* (Boyd et al., 2013). We will use the terminology "ranking the best".

We will now formalise the ranking the best problem, and see how the tools we have developed thus far may be applied to address it.

9.1 Formal definition of ranking the best

Corollary 43 shows the AUC is maximised by any strictly monotone increasing transformation of η, the observation-conditional distribution. Thus, from the perspective of the AUC, the optimal ranked list in bipartite problems involves ordering instances based on their η values. In the ranking the best problem, our goal is to ensure that instances $x \in \mathcal{X}$ for which $\eta(x)$ is large are correctly ordered relative to other instances, potentially at the expense of incorrectly ordering instances $x' \in \mathcal{X}$ for which $\eta(x')$ is small.

Formally, given any $q \in [0,1]$, we call a loss ℓ a *q-RTB loss* (for q-rank-the-best) if the Bayes-optimal scorer for $D = \langle M, \eta \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$ is

$$\mathcal{S}^*(D, \ell) = \{\Psi_q \circ \eta\},$$

where

$$(\forall z \in [0,1])\, \Psi_q(z) \in \begin{cases} \{z\} & \text{if } z \geq q \\ [0, z) & \text{if } z < q, \end{cases} \tag{52}$$

Such a scorer does not demand accurate estimation of η below the fixed threshold q: all that is required is that the ordering is preserved relative to those instances with score bigger than q.

According to Equation 52, any loss for which $\mathcal{S}^*(D, \ell) = \{\Psi \circ \eta\}$ for some invertible Ψ is also a $\Psi^{-1}(q)$-RTB loss for any $q \in [0,1]$. This simply says that if we accurately model *all* ranks, then by definition we accurately model ranks at the head of the list. Indeed, if we could operate on the distribution directly, there would be no tradeoff to be made between accurately modelling any particular portion of the list: η would be recovered exactly, and thus the entire list could be ranked perfectly. The value of a loss that relaxes the modelling requirements for $\eta < q$ arises when we have either finite samples or a misspecified hypothesis class, in which case tradeoffs are necessary.

Having defined the goal of the ranking the best problem, we look to review some performance measures for this task. We then explore alternate risks that have similar characteristics, but are designed using the theory of proper composite losses. We shall begin with a general framework that follows Clémençon and Vayatis (2008); Rudin (2009).

9.2 The (reverse) (ℓ, g)-push framework for ranking the best

Most established performance measures encourage focussing on the head of the ranked list in one of two (related) ways:

(i) ensuring that most negative instances appear below the positive instances, or

(ii) ensuring that most positive instances appear near the head of the list.

The former approach involves ensuring that the false negative rate is small for certain negative instances. The latter approach relies on minimising the following quantity: given an instance $x \in \mathcal{X}$, we define its *normalised rank* under a scorer $s : \mathcal{X} \to \mathbb{R}$ and distribution $D = \langle M, \eta \rangle$ to be the fraction of examples that have a higher score than it:

$$\begin{aligned} \text{NRank}(x; D, s) &\doteq \mathbb{P}_{\mathsf{X} \sim M}[s(\mathsf{X}) > s(x)] + \frac{1}{2} \cdot \mathbb{P}_{\mathsf{X} \sim M}[s(\mathsf{X}) = s(x)] \\ &= \pi \cdot \text{TPR}(s(x); D, s) + (1 - \pi) \cdot \text{FPR}(s(x); D, s). \end{aligned} \tag{53}$$

Small normalised ranks are desired for positive examples, and large ranks for negative examples. Our definition of the normalised rank is simply called the "rank" by Rudin (2009, Section 7).

For each of the above approaches, we consider a general family of risks that can be specialised to yield various performance measures of interest. For approach (i), Rudin (2009); Swamidass et al. (2010) studied a family of risks parameterised by a monotone increasing function[20]. Generalising these proposals to the case of an arbitrary symmetric loss ℓ, and a pair-scorer $s_{\text{Pair}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, we obtain the (ℓ, g)-*push risk*:

$$\text{Push}(s_{\text{Pair}}; D, \ell, g) \doteq \mathop{\mathbb{E}}_{\mathsf{X}' \sim Q} \left[g \left(\mathop{\mathbb{E}}_{\mathsf{X} \sim P} \left[\ell_1(s_{\text{Pair}}(\mathsf{X}, \mathsf{X}')) \right] \right) \right],$$

where $g : \mathbb{R} \to \mathbb{R}_+$ is a monotone increasing function. For decomposable pair-scorers,

$$\text{Push}(\text{Diff}(s); D, \ell, g) = \mathop{\mathbb{E}}_{\mathsf{X}' \sim Q} \left[g \left(\text{FNR}_\ell(s(\mathsf{X}')) \right) \right].$$

Compared to the standard bipartite ℓ-ranking risk (Equation 16), the difference is that the inner expectation is transformed by the function $g(\cdot)$. Intuitively, if $g(\cdot)$ is convex with a sharply increasing slope, the risk penalises high false negative rates, or equally encourages high true positive rates when thresholding around the negatives.

For approach (ii), following Clémençon and Vayatis (2008) and Rudin (2009, Section 7), we have an analogous *reverse*[21] (ℓ, g)-*push risk*, where the order of expectations is reversed, and one uses the normalised rank in place of a rate:

$$\text{RevPush}(s_{\text{Pair}}; D, \ell, g) \doteq \mathop{\mathbb{E}}_{\mathsf{X} \sim P} \left[g \left(\mathop{\mathbb{E}}_{\mathsf{X}' \sim M} \left[\ell_1(s_{\text{Pair}}(\mathsf{X}, \mathsf{X}')) \right] \right) \right],$$

where $g : \mathbb{R} \to \mathbb{R}_+$ is a monotone increasing function, For decomposable pair-scorers,

$$\text{RevPush}(\text{Diff}(s); D, \ell, g) = \mathop{\mathbb{E}}_{\mathsf{X} \sim P} \left[g \left(\text{NRank}_\ell(s(\mathsf{X})) \right) \right],$$

where NRank_ℓ involves replacing the TPR and FPR in Equation 53 with their TPR_ℓ and FPR_ℓ counterparts.

To gain some intuition for these approaches, we show how the AUC is a special case of each.

9.3 The AUC and the (reverse) (ℓ, g)-push risk

We first relate the AUC to the (ℓ, g)-push risk. Recall from Equation 38 that the AUC is

$$\text{AUC}(s; D) = \mathop{\mathbb{E}}_{\mathsf{X}' \sim Q} \left[\text{TPR}(s(\mathsf{X}'); D, s) \right].$$

A high AUC thus means that the negative instances, on average, are placed below the positive instances: were this not the case, then we would achieve only a low true positive rate when thresholding scores around the negatives. Formally, we have

$$\text{AUC}(s; D) = \text{Push}(\text{Diff}(s); D, \ell_{01}, g)$$

20. This family has also been considered, in a different context, by Xie and Priebe (2002).
21. Our use of the word "reverse" is as per Rudin (2009, Section 7).

where $g : x \mapsto 1 - x$.

We next relate the AUC to the reverse (ℓ, g)-push risk. At first glance, the AUC appears to involve a different consideration, as

$$\mathrm{AUC}(s; D) = \mathop{\mathbb{E}}_{\mathsf{X} \sim P} [\mathrm{TNR}(s(\mathsf{X}); D, s)] ,$$

so that the AUC focusses on placing positives ahead of negatives, while when maximising the normalised rank, we also consider the relationship of positives to other positives. (A similar observation is made by Rudin (2009, Section 7), where the false positive rate is called the "reverse height".) However, note that

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathsf{X} \sim P} [\mathrm{NRank}(\mathsf{X}; D, s)] &= \pi \cdot \mathop{\mathbb{E}}_{\mathsf{X} \sim P} [\mathrm{TPR}(s(\mathsf{X}); D, s)] + (1 - \pi) \cdot \mathop{\mathbb{E}}_{\mathsf{X} \sim P} [\mathrm{FPR}(s(\mathsf{X}); D, s)] \\
&= \pi \cdot \mathop{\mathbb{E}}_{\mathsf{X} \sim P} [\mathrm{TPR}(s(\mathsf{X}); D, s)] + (1 - \pi) \cdot (1 - \mathrm{AUC}(s; D)) \text{ by Equation 37} \\
&= \frac{\pi}{2} + (1 - \pi) \cdot (1 - \mathrm{AUC}(s; D)) \text{ following Equation 40} ,
\end{aligned}$$

and so

$$\mathrm{AUC}(s; D) = \frac{2 - \pi}{2 \cdot (1 - \pi)} - \frac{1}{1 - \pi} \cdot \mathop{\mathbb{E}}_{\mathsf{X} \sim P} [\mathrm{NRank}(\mathsf{X}; D, s)] .$$

A high AUC thus means that on average, the positive instances have a small normalised rank i.e. they appear near the head of the list. Formally, we thus have

$$\mathrm{AUC}(s; D) = \mathrm{RevPush}(\mathrm{Diff}(s); D, g)$$

where $g : x \mapsto \frac{2-\pi}{2\cdot(1-\pi)} - \frac{1}{1-\pi} \cdot x$.

9.4 Established performance measures for ranking the best

In both the above interpretations of the AUC, one focusses on average case behaviour. This is manifest in the AUC having a linear dependence on the true positive rate, as well as on the normalised rank. The basic idea of adapting the measure to focus on the head of the list is to consider a suitable nonlinear transformation $g(\cdot)$ in the (reverse) (ℓ, g)-push risk, so as to strongly penalise errors at the head over the tail. We now define some popular measures[22] for ranking the best that do precisely this. Table 9 summarises the measures considered.

9.4.1 PARTIAL AUC

The *partial AUC* (*PAUC*) (McClish, 1989; Dodd and Pepe, 2003; Narasimhan and Agarwal, 2013a) of a scorer only computes the area under the ROC curve for false positive rates between $[a, b] \subseteq [0, 1]$:

$$\begin{aligned}
\mathrm{PAUC}(s; D, a, b) &= \int_a^b \mathrm{TPR}((\mathrm{FPR})^{-1}(\alpha))\, d\alpha \\
&= \mathop{\mathbb{E}}_{\mathsf{X} \sim P} \left[\mathrm{TPR}(s(\mathsf{X})) \cdot [\![a \leq \mathrm{FPR}(s(\mathsf{X})) \leq b]\!] \right] .
\end{aligned}$$

When $a = 0$ and $b \ll 1$, this intuitively focusses only on performance at the head of the ranked list (as this corresponds to thresholds with low false positive rate). This measure is evidently related to the special case of the reverse (ℓ_{01}, g)-push risk for $g : x \mapsto (x \vee a) \wedge b$, where one uses the true positive rate in place of the normalised rank.

22. Most of these measures have their origins in information retrieval. Here, they are typically stated in terms of results for "queries". We effectively treat our labelled samples as the set of results for a single query. Measures that average across multiple queries, which would correspond to a multilabel learning problem, are thus not considered.

Performance measure	Symbol	Definition
Partial AUC	$\mathrm{PAUC}(s; D, b)$	$\mathop{\mathbb{E}}_{\mathsf{X}\sim P}\left[\mathrm{TPR}(s(\mathsf{X})) \cdot [\![\mathrm{FPR}(s(\mathsf{X})) \leq b]\!]\right]$
Average precision	$\mathrm{AP}(s; D)$	$\mathbb{E}_{\mathsf{X}\sim P}\left[\frac{\mathrm{TPR}(s(\mathsf{X}); D, s)}{\mathrm{NRank}(\mathsf{X}; D, s)}\right]$
Discounted cumulative gain	$\mathrm{DCG}(s; D)$	$\mathbb{E}_{\mathsf{X}\sim P}\left[\frac{1}{\lg(1+\mathrm{NRank}(\mathsf{X}; D, s))}\right]$
Average reciprocal rank	$\mathrm{ARR}(s; D)$	$\mathbb{E}_{\mathsf{X}\sim P}\left[\frac{1}{\mathrm{NRank}(\mathsf{X}; D, s)}\right]$
Reciprocal rank	$\mathrm{RR}(s; D)$	$\sup_{x\in\mathrm{supp}(P)} \frac{1}{\mathrm{NRank}(x; D, s)}$
(Negated) p-norm push	$\mathrm{Push}(s; D, \cdot^p)$	$\mathop{\mathbb{E}}_{\mathsf{X}'\sim Q}\left[-\left(\mathrm{FNR}(s(\mathsf{X}'); D, s)\right)^p\right]$
Positives at top	$\mathrm{PTop}(s; D)$	$\inf_{x'\in\mathrm{supp}(Q)} \mathrm{TPR}(s(x'); D, s)$

Table 9: Performance measures for ranking the best. For each measure, larger values are desirable.

9.4.2 AVERAGE PRECISION

Our next measure relies on the following two quantities.

Definition 50 *Given any distribution $D \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ and scorer $s : \mathcal{X} \to \mathbb{R}$, define the precision and recall at a threshold $t \in \mathbb{R} \cup \{\pm\infty\}$ to be*

$$\mathrm{Prec}(t; D, s) \doteq \mathbb{P}[\mathsf{Y} = 1 | s(\mathsf{X}) > t]$$
$$\mathrm{Rec}(t; D, s) \doteq \mathrm{TPR}(t; D, s) = \mathbb{P}[s(\mathsf{X}) > t | \mathsf{Y} = 1] + \frac{1}{2} \cdot \mathbb{P}[s(\mathsf{X}) = t | \mathsf{Y} = 1].$$

When the scorer and distribution are clear from context, we shall drop the dependence on them and simply write $\mathrm{Prec}(t)$, $\mathrm{Rec}(t)$.

The precision may be related to the more familiar rates introduced earlier: if the distribution of scores has no discrete components, then by Bayes' rule,

$$\begin{aligned}\mathrm{Prec}(t; D, s) &= \frac{\mathbb{P}[s(\mathsf{X}) > t | \mathsf{Y} = 1] \cdot \mathbb{P}[\mathsf{Y} = 1]}{\mathbb{P}[s(\mathsf{X}) > t]} \\ &= \frac{\pi \cdot \mathrm{TPR}(t)}{\pi \cdot \mathrm{TPR}(t) + (1 - \pi) \cdot \mathrm{FPR}(t)} \\ &= \left(1 + \frac{1 - \pi}{\pi} \cdot \frac{\mathrm{FPR}(t)}{\mathrm{TPR}(t)}\right)^{-1}.\end{aligned} \tag{54}$$

Note that if we use as threshold $t = s(x)$ for some $x \in \mathcal{X}$, then the denominator of Equation 54 is nothing but $\mathrm{NRank}(t)$.

We now define the *average precision (AP)* of a scorer s (Yue et al., 2007; Chakrabarti et al., 2008; Agarwal, 2011; Boyd et al., 2012) to be the average of the precisions obtained using the scores of positive examples as thresholds:

$$\begin{aligned}\mathrm{AP}(s; D) &\doteq \mathbb{E}_{\mathsf{X}\sim P}\left[\mathrm{Prec}(s(\mathsf{X}); D, s)\right] \\ &= \pi \cdot \mathbb{E}_{\mathsf{X}\sim P}\left[\frac{\mathrm{TPR}(s(\mathsf{X}); D, s)}{\mathrm{NRank}(\mathsf{X}; D, s)}\right],\end{aligned} \tag{55}$$

where the second equation follows from Equations 53 and 54.

The average precision can be shown to favour accuracy at the head of the list more than the AUC (Yue et al., 2007). Intuitively, this is because when there is a spurious negative example high in the ranked list, it will substantially affect the precision of the nearby positive examples. This may also be seen through Equation 55: we encourage very low normalised ranks for the positive examples, which corresponds to placing them at the very top of the list. Compared to the AUC, placing a few positives at the very top gives a greater gain than placing many positives only roughly near the top.

Further intuition for the average precision can be gained by considering the precision-recall curve, a complement to the ROC curve.

Definition 51 *Given any distribution* $D \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ *and scorer* $s : \mathcal{X} \to \mathbb{R}$*, the precision-recall curve is defined by the parametric representation*

$$\mathrm{PR}(s; D) \doteq \{(\mathrm{Rec}(t; D, s), \mathrm{Prec}(t; D, s)) : t \in \mathbb{R} \cup \{\pm\infty\}\} \subseteq [0, 1]^2.$$

The area under the precision recall curve (AUPRC) *of* s *is the area under the curve* $\mathrm{PR}(s; D)$ *(Boyd et al., 2013):*

$$\mathrm{AUPRC}(s; D) \doteq \int_0^1 \mathrm{Prec}(\mathrm{Rec}^{-1}(\alpha))\, d\alpha.$$

We immediately see that compared to the ROC curve, the PR curve depends on the base rate π. We also see that like the ROC curve, the curve can be computed from the TPR and FPR. Following the Neyman-Pearson analysis (Corollary 16), we can conclude that the area under the PR curve is optimized by any strictly monotone increasing transform of $\eta(x)$ (Clémençon and Vayatis, 2009a).

In fact, the average precision is exactly equal to the AUPRC.

Lemma 52 ((Boyd et al., 2013)) *Given any* $D \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ *and scorer* $s : \mathcal{X} \to \mathbb{R}$ *with differentiable ROC curve and invertible rates,*

$$\mathrm{AUPRC}(s; D) = \mathrm{AP}(s; D).$$

Proof By definition,

$$\begin{aligned}
\mathrm{AUPRC}(s; D) &= \int_0^1 \mathrm{Prec}(\mathrm{Rec}^{-1}(\alpha))\, d\alpha \\
&= -\int_{-\infty}^{\infty} \mathrm{TPR}'(t) \cdot \mathrm{Prec}(t)\, dt \text{ using } \alpha = \mathrm{TPR}(t) \\
&= \int_{-\infty}^{\infty} p_S(t) \cdot \mathrm{Prec}(t)\, dt \text{ by Equation 22} \\
&= \mathop{\mathbb{E}}_{\mathsf{X}\sim P}\left[\int_{-\infty}^{\infty} \delta_{s(\mathsf{X})}(t) \cdot \mathrm{Prec}(t)\, dt\right] \\
&= \mathop{\mathbb{E}}_{\mathsf{X}\sim P}\left[\mathrm{Prec}(s(\mathsf{X}))\right].
\end{aligned}$$

■

9.4.3 DISCOUNTED CUMULATIVE GAIN

The *discounted cumulative gain* (*DCG*) of a scorer s (Järvelin and Kekäläinen, 2002; Agarwal, 2011; Boyd et al., 2012) is the average of the inverse logarithm of the normalised rank for all positive examples:

$$\mathrm{DCG}(s; D) \doteq \mathbb{E}_{\mathsf{X}\sim P}\left[\frac{1}{\lg(1 + \mathrm{NRank}(\mathsf{X}; D, s))}\right].$$

Compared to average precision (Equation 55), the DCG applies a nonlinear decay on the effect of lower ranked positives. It is evident that the DCG is a special case of the reverse (ℓ_{01}, g)-push risk with $g : x \mapsto 1/\log(1+x)$. It may also be seen as a limiting case of the family of risks with $g_p : x \mapsto (1/p)/((1+x)^{1/p}-1)$ as $p \to \infty$.

9.4.4 AVERAGE RECIPROCAL RANK

The *average reciprocal rank (ARR)* of a scorer s (Rudin, 2009, Section 7) is the inverse of the harmonic mean of the normalised ranks for all positive examples:

$$\text{ARR}(s; D) = \mathbb{E}_{\mathsf{X} \sim P}\left[\frac{1}{\text{NRank}(\mathsf{X}; D, s)}\right].$$

This measure encourages small normalised rank values for the positives, meaning that spurious negatives near the head of the list will adversely affect the scores for several examples. Compared to the average precision (Equation 55), one does not additionally weigh this inverse rank by the true positive rate. It is evident that the ARR is a special case of the reverse (ℓ_{01}, g)-push risk with $g : x \mapsto 1/x$.

9.4.5 RECIPROCAL RANK

The *reciprocal rank (RR)* of a scorer s (Voorhees, 2001; Chakrabarti et al., 2008) is the inverse of the rank of the top positive:

$$\text{RR}(s; D) \doteq \sup_{x \in \text{supp}(P)} \frac{1}{\text{NRank}(x; D, s)}.$$

where $\text{supp}(\cdot)$ denotes the support of a distribution. This measure directly encourages the first element of the ranked list to be a positive. Compared to average precision (Equation 55), roughly, we replace the average performance over all positives with simply the performance of the best positive. The RR can be seen as a limiting case of a family of reverse (ℓ_{01}, g)-push risks with $g = g_p : x \mapsto 1/x^p$ as $p \to \infty$.

Compared to the RR, the ARR considers the ranks of *all* positives, not just the top one. This intuitively makes the ARR more suitable when one is interested not just at the very first element of the list, but rather on the first k elements for some small constant k.

9.4.6 THE p-NORM PUSH

Rudin (2009) provides a detailed study of the *p-norm push*, which the (ℓ, g)-push risk for the choice $g : x \mapsto x^p$ for $p \in [1, \infty)$ and symmetric ℓ, leading to the *p-norm push* risk:

$$\begin{aligned}\text{Push}(\text{Diff}(s); D, \ell, \cdot^p) &= \mathop{\mathbb{E}}_{\mathsf{X}' \sim Q}\left[\left(\text{FNR}_\ell(s(\mathsf{X}'))\right)^p\right] \\ &= \mathop{\mathbb{E}}_{\mathsf{X}' \sim Q}\left[\left(\mathop{\mathbb{E}}_{\mathsf{X} \sim P}\left[\ell_1(s(\mathsf{X}) - s(\mathsf{X}'))\right]\right)^p\right]. \end{aligned} \tag{56}$$

By increasing the value of p, one strongly penalises high false negative rates.

9.4.7 POSITIVES AT THE TOP

The *fraction of positives at the top* (*PTop*) of a scorer s (Agarwal, 2011; Boyd et al., 2012) is typically defined on an empirical sample $\mathcal{D} = \{(x_i, 1)\}_{i=1}^n \cup \{(x_j, -1)\}_{j=1}^m$ as the number of positive instances ranked above the highest negative instance, or equally, the minimum over all negative instances of the number of positives ranked above that instance,

$$\text{PTop}(s; \mathcal{D}) \doteq \min_{1 \le j \le m} \sum_{i=1}^{n} \ell_{01}(s(x_j) - s(x_i)).$$

Li et al. (2014) provided an efficient algorithm to optimise surrogates to this measure. Evidently, its population counterpart is

$$\mathrm{PTop}(s; D) \doteq \inf_{x' \in \mathrm{supp}(Q)} \mathrm{TPR}(s(x'); D, s),$$

with negation

$$1 - \mathrm{PTop}(s; D) = \sup_{x' \in \mathrm{supp}(Q)} \mathrm{FNR}(s(x'); D, s). \tag{57}$$

Compared to the AUC (Equation 38), which looks to make the *average* rank of all negative instances small, the PTop looks to make the *worst possible* rank over all negative instances small. The PTop can be related to the p-norm push risk, as

$$\lim_{p \to \infty} \mathrm{Push}(s; D, \ell, \cdot^p)^{1/p} = \sup_{x' \in \mathrm{supp}(Q)} \mathrm{FNR}_\ell(s(x')).$$

For the case of ℓ_{01}, this is exactly the negation of the positives at the top measure (Equation 57).

9.5 Bayes-optimal scorers for the (ℓ, g)-push risk

Having introduced a number of performance measures, we now study why suitable choices of g for the (ℓ, g)-push risk can be seen to focus attention at the head of the list. This is done by analysing the Bayes-optimal scorers for this family of risks, and seeing how they align with Equation 52. Specifically, we aim to determine the Bayes-optimal pair and univariate scorers for the (ℓ, g)-push risk, and study them in light of Equation 52. Unlike bipartite ranking, the risk in this case cannot (obviously) be expressed as a classification risk over pairs of instances; therefore, we separately consider the optimal pair- and univariate-scorers,

$$\mathcal{S}^{\mathrm{pair},*}_{\mathrm{push}}(D, \ell, g) = \operatorname*{Argmin}_{s_{\mathrm{Pair}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}} \mathrm{Push}(s_{\mathrm{Pair}}; D, \ell, g)$$
$$\mathcal{S}^{*}_{\mathrm{push}}(D, \ell, g) = \operatorname*{Argmin}_{s : \mathcal{X} \to \mathbb{R}} \mathrm{Push}(\mathrm{Diff}(s); D, \ell, g).$$

We first analyse the case of pair-scorers, and then proceed to univariate scorers. While most of our analysis is for general ℓ and g, we shall find the p-norm push risk of Equation 56 to be particularly amenable to analysis when combined with the exponential loss.

9.5.1 BAYES-OPTIMAL PAIR-SCORERS

As with the standard bipartite risk, determining the Bayes-optimal scorer for the (ℓ, g) push is challenging due to the implicit restricted function class $\mathcal{S}_{\mathrm{Decomp}}$. In fact, this is difficult even for the pair-scorer case: the (ℓ, g) push risk is not easily expressible in terms of a conditional risk. Thus, we explicitly compute the derivative of the risk, as in the proof of Proposition 47. We end up with the following distribution-dependent transformation of η_{Pair} as our optimal scorer.

Proposition 53 *Given any $D = \langle M, \eta \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$, a differentiable function $g : \mathbb{R} \to \mathbb{R}$, and a differentiable strictly proper composite loss ℓ with link function Ψ, if ℓ'_1, ℓ'_{-1} are bounded or $\mathcal{X}$ is finite,*

$$\mathcal{S}^{\mathrm{pair},*}_{\mathrm{push}}(D, \ell, g) = \left\{ s^*_{\mathrm{Pair}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R} : s^*_{\mathrm{Pair}} = \Psi \circ \sigma \circ (\mathrm{Diff}(\sigma^{-1} \circ \eta) - G(D, s^*_{\mathrm{Pair}})) \right\}, \tag{58}$$

where

$$G(x, x'; D, s_{\mathrm{Pair}}) \doteq \log \frac{g'\left(F(x; D, s_{\mathrm{Pair}})\right)}{g'\left(F(x'; D, s_{\mathrm{Pair}})\right)}$$
$$F(x; D, s_{\mathrm{Pair}}) \doteq \mathop{\mathbb{E}}_{\mathsf{X} \sim P} \left[\frac{\ell_1(s_{\mathrm{Pair}}(\mathsf{X}, x)) + \ell_{-1}(s_{\mathrm{Pair}}(x, \mathsf{X}))}{2} \right].$$

Proof First, in the notation above,

$$\text{Push}(s_{\text{Pair}}; D, \ell, g) = \mathop{\mathbb{E}}_{\mathsf{X}' \sim Q} \left[g\left(F(\mathsf{X}'; D, s_{\text{Pair}}) \right) \right].$$

For fixed D, let $\mathcal{S}(D)$ denote the space of all Lebesgue-measurable pair-scorers $s_{\text{Pair}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, with addition and scalar multiplication defined pointwise, such that $\text{Push}(s_{\text{Pair}}; D, \ell, g) < \infty$. As before, we consider the Gâteaux variation of the functional. Pick any $s_{\text{Pair}}, t_{\text{Pair}} \in \mathcal{S}(D)$. For any $\epsilon > 0$, define

$$\begin{aligned} R(\epsilon; s_{\text{Pair}}, t_{\text{Pair}}) &= \text{Push}(s_{\text{Pair}} + \epsilon \cdot t_{\text{Pair}}; D, \ell) \\ &= \mathop{\mathbb{E}}_{\mathsf{X}' \sim Q} \left[g\left(F(\mathsf{X}'; D, s_{\text{Pair}} + \epsilon \cdot t_{\text{Pair}}) \right) \right]. \end{aligned}$$

For simplicity, in the following we shall not explicitly write the dependence of F and G on D, s_{Pair}. Now consider

$$\begin{aligned} R'(0; s_{\text{Pair}}, t_{\text{Pair}}) &= \mathop{\mathbb{E}}_{\mathsf{X}' \sim Q} \left[g'\left(F(\mathsf{X}') \right) \cdot \mathop{\mathbb{E}}_{\mathsf{X} \sim P} \left[t_{\text{Pair}}(\mathsf{X}, \mathsf{X}') \cdot \frac{\ell'_1(s_{\text{Pair}}(\mathsf{X}, \mathsf{X}'))}{2} + \right. \right. \\ &\qquad \left. \left. t_{\text{Pair}}(\mathsf{X}', \mathsf{X}) \cdot \frac{\ell'_{-1}(s_{\text{Pair}}(\mathsf{X}', \mathsf{X}))}{2} \right] \right] \\ &= \frac{1}{2} \int_{\mathcal{X} \times \mathcal{X}} t_{\text{Pair}}(x, x') \cdot \left(p(x) q(x') \cdot g'(F(x')) \cdot \ell'_1(s_{\text{Pair}}(x, x')) + \right. \\ &\qquad \left. p(x') q(x) \cdot g'(F(x)) \cdot \ell'_1(s_{\text{Pair}}(x', x)) \right) dx \, dx', \end{aligned}$$

where as in the proof of Proposition 47, the interchange of derivative and expectation is justified when $\mathcal{X}$ is finite, or when the derivatives ℓ'_1, ℓ'_{-1} are bounded.

For the optimal pair-scorer s^*_{Pair}, the derivative must be zero for every t_{Pair}. A sufficient condition for this to hold is that the second term in the integrand is zero for (almost) every $x, x' \in \mathcal{X}$.

Now, since ℓ is strictly proper composite, for any $\eta \in [0, 1]$, the solution to

$$\eta \cdot \ell'_1(s) + (1 - \eta) \cdot \ell'_{-1}(s) = 0$$

is $s = \Psi(\eta)$, by virtue of the above being the derivative of the conditional risk. Thus, the solution to

$$\frac{a}{a+b} \cdot \ell'_1(s) + \frac{b}{a+b} \cdot \ell'_{-1}(s) = 0$$

for $a, b > 0$ is $s = \Psi(a/(a+b)) = \Psi(\sigma(\log(a/b)))$. Letting

$$\begin{aligned} a &\doteq g'\left(F(x') \right) \cdot p(x) \cdot q(x') \\ b &\doteq g'\left(F(x) \right) \cdot q(x) \cdot p(x'), \end{aligned}$$

the optimal pair-scorer is, for every $x, x' \in \mathcal{X}$,

$$\begin{aligned} s^*_{\text{Pair}}(x, x') &= \Psi \circ \sigma \circ \log \frac{p(x) \cdot q(x') \cdot g'(F(x'))}{p(x') \cdot q(x) \cdot g'(F(x))} \\ &= \Psi \circ \sigma \circ \left(\sigma^{-1}(\eta(x)) - \sigma^{-1}(\eta(x')) - G(D, s^*_{\text{Pair}}) \right), \end{aligned}$$

where the second line is since

$$\frac{p(x)}{q(x)} = \frac{\eta(x)}{1 - \eta(x')} \cdot \frac{1 - \pi}{\pi}.$$

Thus,

$$s^*_{\text{Pair}} = \Psi \circ \sigma \circ \left(\text{Diff}(\sigma^{-1} \circ \eta) - G(D, s^*_{\text{Pair}}) \right).$$

The result follows by dividing through by the numerator. ■

Requiring differentiability of the loss means that we cannot compute the optimal solution for ℓ_{01}. However, we can compute the Bayes-optimal pair-scorers for a sequence of proper composite losses that approach ℓ_{01}. Consider a sequence of losses $\{\ell^{(n)}\}_{n\in\mathbb{N}}$ with corresponding links $\{\Psi^{(n)}\}_{n\in\mathbb{N}}$, with $(\forall v \in \mathbb{R})\ \lim_{n\to\infty}(\Psi^{(n)})^{-1}(v) = [\![v > 0]\!]$. This suggests the optimal scorer of

$$\mathcal{S}^{\text{pair},*}_{\text{push}}(D, \ell_{01}, g) = \left\{ s^*_{\text{Pair}} : \mathcal{X}\times\mathcal{X} \to \mathbb{R} : \text{sign}(s^*_{\text{Pair}}) = \text{sign}(\text{Diff}(\sigma^{-1}\circ\eta) - G(D, s^*_{\text{Pair}})) \right\},$$

but we defer a formal proof to future work.

When $g : x \mapsto x$, which corresponds to the standard ℓ-bipartite ranking risk, the term G above is $\equiv 0$ and so $s^*_{\text{Pair}} = \Psi\circ\eta_{\text{Pair}}$ as expected. For general (ℓ, g), however, it is unclear how to simplify the term G any further. In general, s^*_{Pair} appears to be a strictly monotone transform of η_{Pair}, where the transform is distribution dependent. However, surprisingly, for the special case of ℓ being the exponential loss and $g : x \mapsto x^p$, the optimal scorer is explicitly determinable as a simple transform of the conditional probability.

Proposition 54 *Pick any distribution $D = \langle M, \eta\rangle \in \Delta_{\mathcal{X}\times\{\pm1\}}$. Let $\ell(y, v) = e^{-yv}$ be the exponential loss and $g : x \mapsto x^p$ for any $p \geq 1$. Then, the optimal pair-scorer s^*_{Pair} for the (ℓ, g)-push bipartite ranking risk is*

$$s^*_{\text{Pair}} = \frac{1}{p+1}\cdot \text{Diff}(\sigma^{-1}\circ\eta).$$

Proof We establish this by verifying that $s_{\text{Pair}} = \frac{1}{p+1}\text{Diff}(\sigma^{-1}\circ\eta)$ satisfies the implicit equation in Equation 58. We begin with the term $F(x; D, s_{\text{Pair}})$ as defined in Proposition 53. Plugging in $g : x \mapsto x^p$ and

$$s_{\text{Pair}} = \frac{1}{p+1}\cdot\sigma^{-1}\circ\eta_{\text{Pair}} = \frac{1}{p+1}\cdot\text{Diff}(\sigma^{-1}\circ\eta),$$

we get

$$\begin{aligned}
(\forall x \in \mathcal{X})\ F(x; D, s_{\text{Pair}}) &= \mathop{\mathbb{E}}_{\mathsf{X}\sim P}\left[\frac{\ell_1(s_{\text{Pair}}(\mathsf{X}, x)) + \ell_{-1}(s_{\text{Pair}}(x, \mathsf{X}))}{2}\right] \\
&= \mathop{\mathbb{E}}_{\mathsf{X}\sim P}\left[\frac{e^{-s_{\text{Pair}}(\mathsf{X},x)} + e^{s_{\text{Pair}}(x,\mathsf{X})}}{2}\right] \\
&= \frac{1}{2}\mathop{\mathbb{E}}_{\mathsf{X}\sim P}\left[\left(\frac{\eta_{\text{Pair}}(\mathsf{X}, x)}{1-\eta_{\text{Pair}}(\mathsf{X}, x)}\right)^{-1/(p+1)} + \left(\frac{\eta_{\text{Pair}}(x, \mathsf{X})}{1-\eta_{\text{Pair}}(x, \mathsf{X})}\right)^{1/(p+1)}\right] \\
&= \mathop{\mathbb{E}}_{\mathsf{X}\sim P}\left[\exp((\sigma^{-1}(\eta(x)) - \sigma^{-1}(\eta(\mathsf{X})))/(p+1))\right] \\
&= \exp(\sigma^{-1}(\eta(x))/(p+1))\cdot\mathop{\mathbb{E}}_{\mathsf{X}\sim P}\left[\exp(-\sigma^{-1}(\eta(\mathsf{X})))/(p+1))\right],
\end{aligned}$$

where crucially the dependence on η is separated from the dependence on the rest of the distribution.

Thus, for $g : x \mapsto x^p$,

$$(\forall x, x' \in \mathcal{X})\ \frac{g'\left(F(x; D, s_{\text{Pair}})\right)}{g'\left(F(x'; D, s_{\text{Pair}})\right)} = \frac{\exp(\sigma^{-1}(\eta(x))\cdot(p-1)/(p+1))}{\exp(\sigma^{-1}(\eta(x'))\cdot(p-1)/(p+1))}$$

with the result now a simple function of η, and

$$(\forall x, x' \in \mathcal{X})\ \log\frac{g'\left(F(x; D, s_{\text{Pair}})\right)}{g'\left(F(x'; D, s_{\text{Pair}})\right)} = \frac{p-1}{p+1}\cdot(\sigma^{-1}(\eta(x)) - \sigma^{-1}(\eta(x')).$$

Now recall that the link function for exponential loss is $\Psi = \frac{1}{2}\sigma^{-1}$. Plugging the above into the right hand side of Equation 58, we get

$$\begin{aligned}\Psi \circ \sigma \circ (\mathrm{Diff}(\sigma^{-1} \circ \eta) - G(D, s^*_{\mathrm{Pair}})) &= \left(\frac{1}{2} - \frac{p-1}{2(p+1)}\right) \cdot \mathrm{Diff}(\sigma^{-1} \circ \eta) \\ &= \frac{1}{p+1} \cdot \mathrm{Diff}(\sigma^{-1} \circ \eta) \\ &= s_{\mathrm{Pair}}.\end{aligned}$$

Therefore $s_{\mathrm{Pair}} = \frac{1}{p+1}\mathrm{Diff}(\sigma^{-1} \circ \eta)$ satisfies the implicit equation of Proposition 53, and hence must be an optimal pair-scorer for exponential loss. ■

To see why exponential loss simplifies matters, we note that the risk can be decomposed into

$$\mathrm{Push}(\mathrm{Diff}(s); D, \exp, \cdot^p) = \left(\mathop{\mathbb{E}}_{\mathsf{X} \sim P}\left[e^{-s(\mathsf{X})}\right]\right)^p \cdot \left(\mathop{\mathbb{E}}_{\mathsf{X}' \sim Q}\left[e^{p \cdot s(\mathsf{X}')}\right]\right).$$

This decomposition into the product of two expectations simplifies the derivatives considerably. In fact, an alternate strategy to determine the minimisers of the risk is to consider

$$\operatorname*{argmax}_{s\,:\,\mathcal{X} \to \mathbb{R}} \mathop{\mathbb{E}}_{\mathsf{X}' \sim Q}\left[e^{p \cdot s(\mathsf{X}')}\right] \;:\; \left(\mathop{\mathbb{E}}_{\mathsf{X} \sim P}\left[e^{-s(\mathsf{X})}\right]\right)^p \leq C;$$

this is reminiscent of the Neyman-Pearson approach to arguing for the optimal scorers for the AUC (which incidentally is the strategy we shall employ for proving Proposition 56).

As with Proposition 47, we suspect the finiteness assumption on $\mathcal{X}$ can be dropped, although we have been unsuccessful in establishing this. Nonetheless, for this special case, the optimal scorer can be expressed as $\frac{2}{p+1} \cdot \Psi \circ \eta_{\mathrm{Pair}}$, where Ψ is the link function corresponding to exponential loss; comparing this to the optimal pair-scorer for the standard bipartite risk (Equation 50), we see that the effect of the function $g : x \mapsto x^p$ is equivalent to slightly transforming the loss ℓ; we will explore this more in the next section.

For other losses, the optimal pair-scorer appears to be a genuinely distribution specific transformation of η_{Pair}, as we illustrate in Appendix I.

9.5.2 BAYES-OPTIMAL UNIVARIATE SCORERS

We now turn attention to computing $\mathcal{S}^*_{\mathrm{push}}(D, \ell, g)$. For ℓ_{01}, we were unsuccessful in computing the optimal pair-scorer; nonetheless, a different technique lets us establish the optimal univariate scorers. The basic observation is that the (ℓ_{01}, g)-push risk can be interpreted as the area under the parametric curve

$$\{(\mathrm{FPR}(t; D, s), g(\mathrm{FNR}(t; D, s))) : t \in \mathbb{R}\},$$

which, compared to the ROC curve $\mathrm{ROC}(s; D)$, transforms the true positive rates (or, equivalently, one minus the true negative rates) at each corresponding false positive rate. By manipulating the choice of g, the area under this curve can thus be focus more attention on certain ranges of false negative rates. The equivalence is formalised below.

Proposition 55 *Given any $D \in \Delta_{\mathcal{X} \times \{\pm 1\}}$, $g : \mathbb{R} \to \mathbb{R}$, and scorer $s : \mathcal{X} \to \mathbb{R}$ with differentiable ROC curve and invertible rates,*

$$\mathrm{Push}(\mathrm{Diff}(s); D, \ell_{01}, g) = \int_0^1 g\left(\mathrm{FNR}(\mathrm{FPR}^{-1}(\alpha))\right) \, d\alpha.$$

Proof The proof follows how we established the 0-1 bipartite risk to an area under the curve (Proposition 21):

$$\begin{aligned}
\text{Push}(\text{Diff}(s); D, \ell_{01}, g) &= \mathop{\mathbb{E}}_{X' \sim Q} \left[g\left(\text{FNR}(s(X'))\right) \right] \\
&= \mathop{\mathbb{E}}_{X' \sim Q} \left[\int_{-\infty}^{\infty} \delta_{s(X')}(t) \cdot g\left(\text{FNR}(t)\right) \, dt \right] \\
&= \int_{-\infty}^{\infty} \mathop{\mathbb{E}}_{X' \sim Q} \left[\delta_{s(X')}(t) \cdot g\left(\text{FNR}(t)\right) \right] \, dt \\
&= \int_{-\infty}^{\infty} q_S(t) \cdot g\left(\text{FNR}(t)\right) \, dt \\
&= \int_{-\infty}^{\infty} -\text{FPR}'(t) \cdot g\left(\text{FNR}(t)\right) \, dt \text{ by Equation 22} \\
&= \int_{0}^{1} g\left(\text{FNR}(\text{FPR}^{-1}(\alpha))\right) \, d\alpha.
\end{aligned}$$

■

We can now establish the Bayes-optimal univariate scorers. (A similar result for the case of the reverse (ℓ, g)-push risk was shown in Clémençon and Vayatis (2008, Proposition 7).)

Proposition 56 *Let g be a nonnegative, monotone increasing function. Given any $D = \langle M, \eta \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$,*

$$\phi \circ \eta \in \mathcal{S}^*_{\text{push}}(D, \ell_{01}, g),$$

for any strictly monotone increasing $\phi : [0,1] \to \mathbb{R}$.

Proof Recall from Proposition 55 that the p-norm risk for the case of 0-1 loss is simply an area under the parametric curve

$$\left\{ (\text{FPR}(t), g(\text{FNR}(t))) : t \in \mathbb{R} \cup \{\pm\infty\} \right\}.$$

Following the Neyman-Pearson approach to ROC maximisation (Proposition 76), maximisation of the 0-1 risk is thus equivalent to solving, for each $\alpha \in [0,1]$

$$\mathop{\text{Argmin}}_{s\,:\,\mathcal{X} \to \mathbb{R}, t \in \mathbb{R} \cup \{\pm\infty\}} g(\text{FNR}(t; D, s)) \text{ subject to } \text{FPR}(t; D, s) \leq \alpha.$$

Since g is a monotone increasing function, it preserves the optimal solution of the case of $g(x) = x$ (although potentially introducing new ones), which is the standard Neyman-Pearson problem. This means that for monotone increasing g, one family of optimal solutions is given by $s^* = \phi \circ \eta$, where ϕ is strictly monotone increasing. ■

Proposition 56 says that the (ℓ_{01}, g)-push objective is optimised by accurately recovering the entire ranked list. However, while they share the same optimal solution, the ordering over scorers induced by the (ℓ_{01}, g)-push risk is different from that induced by the standard bipartite ranking risk. This means that under misspecification or with finite samples, one will likely choose a different scorer by optimising the (ℓ_{01}, g)-push objective rather than the standard bipartite ranking objective. The examples in Rudin (2009) indicate that in many such cases, the solutions of the (ℓ_{01}, g)-push objective are superior to those of bipartite ranking at the head of the list.

The above trick does not work when we use a general proper composite loss ℓ, as we need to analyse a generalised Neyman-Pearson problem. However, for exponential loss and $g : x \mapsto x^p$, we can use the results of the previous section.

Proposition 57 *Pick any* $D = \langle M, \eta \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$. *Let* $\ell = \ell_{\exp}$ *and* $p > 0$. *Then, if* $\mathcal{X}$ *is finite,*

$$\mathcal{S}^*_{\text{push}}(D, \ell_{\exp}, \cdot^p) = \left\{ \frac{1}{p+1} \cdot (\sigma^{-1} \circ \eta) + b \,:\, b \in \mathbb{R} \right\}.$$

Proof By Proposition 54, the unique optimal pair-scorer is $s^*_{\text{Pair}} = \frac{1}{p+1} \cdot \text{Diff}(\sigma^{-1} \circ \eta) = \text{Diff}\left(\frac{1}{p+1}(\sigma^{-1} \circ \eta)\right)$, which is decomposable. Corollary 40 may be adapted here to argue that any optimal univariate scorer s^* must satisfy $s^*_{\text{Pair}} = \text{Diff}(s^*)$, and so $s^* = \frac{1}{p+1} \cdot (\sigma^{-1} \circ \eta) + b$ for some $b \in \mathbb{R}$. ■

For other losses, the optimal univariate scorer again appears to be distribution specific, as we illustrate in Appendix J.

As before, the Bayes-optimal scorers for the p-norm push are closely related to those for appropriate proper composite losses (namely, those with link functions given by $\Psi = \frac{1}{p+1} \cdot \sigma$). We now study how the theory of proper composite losses suggests a recipe for constructing a family of alternate losses suitable for the ranking the best tasks.

9.6 Proper composite losses for ranking the best

Having studied the Bayes-optimality properties of the p-norm push, we now examine what this implies about the design of alternate proper composite losses for ranking the best. As shall be made precise, the p-norm push can be understood in terms of a suitable weight function over misclassification costs.

9.6.1 A WEIGHT FUNCTION PERSPECTIVE OF THE p-NORM PUSH

From Proposition 57, we see that changing p results in a scaling of the link function Ψ that is composed with η. Thus, the p-norm push has equivalent Bayes-optimal solutions, up to translation, as any strictly proper composite loss with the same link function $\Psi^{(p)} = \frac{1}{p+1} \cdot \sigma^{-1}$. One might then hope to understand the p-norm push risk by considering the risks corresponding to a family of proper composite losses $\{\ell^{(p)}\}_{p \in \mathcal{P}}$, where each member of the family comprises some fixed proper loss λ composed with an appropriately scaled sigmoidal link $\Psi^{(p)}$. However, for any $p > 0$, the resulting proper composite loss is

$$\ell^{(p)}(y, v) = \lambda(y, \Psi^{(p)}(v)) = \lambda(y, \sigma((p+1) \cdot v)) = \ell^{(0)}(y, (p+1)v).$$

That is, changing p simply scales the prediction space, and has no real impact on learning. This means that even on a finite sample, and with a restricted function class, the family of proper composite losses given by $\{\ell^{(p)}\}_{p \in \mathcal{P}}$ will have risks whose optimal solutions that are scalings of one another.

As with the ℓ_{01} case, this is not surprising. It merely indicates that the p-norm push risk must be understood in terms of its behaviour under a restricted function class or finite sample. Doing so requires that one move away from Bayes-optimal scorers, which assume access to infinite samples and an unrestricted function class. Our standard analysis based on the conditional risk thus cannot be applied.

Remarkably, it is possible to show that the p-norm push risk is equivalent to a specific proper composite risk even when minimising over a linear function class: Ertekin and Rudin (2011, Theorem 1) shows that for a linear function class, the p-norm push risk with exponential loss is equivalent to the proper composite risk corresponding to the *p-classification loss*, defined by

$$(\forall v \in \mathbb{R})\, \ell_{\text{pcl}}(v; p) \doteq \left(\frac{1}{p} \cdot e^{vp}, e^{-v} \right). \tag{59}$$

Interestingly, this loss is proper composite.

Lemma 58 *For any $p > 0$, let $\ell = \ell_{\text{pcl}}(\cdot; p)$ be the p-classification loss of Equation 59. Then, $\ell \in \mathcal{L}_{\text{SPC}}(\Psi^{(p)})$, where $(\Psi^{(p)})^{-1} : v \mapsto \sigma((p+1) \cdot v)$. Further, the underlying proper loss $\lambda_{\text{pcl}}(\cdot; p)$ is*

$$(\forall u \in [0,1])\ \lambda_{\text{pcl}}(u; p) = \left(\frac{1}{p} \cdot \left(\frac{u}{1-u} \right)^{1-\frac{1}{p+1}}, \left(\frac{1-u}{u} \right)^{\frac{1}{p+1}} \right).$$

Proof We can check that for $\Psi = \Psi^{(p)}$,

$$\begin{aligned}(\forall v \in \mathbb{R})\ \Psi^{-1}(v) &= \frac{1}{1 - \frac{\ell'_1(v)}{\ell'_{-1}(v)}} \\ &= \frac{1}{1 + e^{-(p+1) \cdot v}} \\ &= \sigma((p+1) \cdot v),\end{aligned}$$

which is invertible, thus guaranteeing that ℓ is proper composite.

It may be checked that the underlying proper loss is

$$\begin{aligned}(\forall u \in [0,1])\ \lambda_{\text{pcl}}(u; p) &= \ell_{\text{pcl}}(\Psi^{(p)}(u); p) \\ &= \left(\frac{1}{p} \cdot \left(\frac{u}{1-u} \right)^{1-\frac{1}{p+1}}, \left(\frac{1-u}{u} \right)^{\frac{1}{p+1}} \right).\end{aligned}$$

■

Given the the loss ℓ_{pcl} is proper composite, the agreement of the p-norm and p-classification risk minimisers is trivial in the unrestricted function class setting; however, it is not obvious in the linear class setting.

The equivalence to p-classification is valuable, since we can analyse the proper composite loss to understand how it focusses accuracy at the head of the ranked list. We will do this by considering the corresponding weight function for the proper loss $\lambda = \lambda_{\text{pcl}}(\cdot; p)$,

$$\begin{aligned}(\forall c \in (0,1))\ w_{\text{pcl}}(c; p) &= -\frac{\lambda'_1(c)}{1-c} \text{ (by (Reid and Williamson, 2010, Theorem 1))} \\ &= \frac{1}{p+1} \cdot \frac{1}{c^{1+\frac{1}{p+1}} \cdot (1-c)^{2-\frac{1}{p+1}}}.\end{aligned}$$

This is a generalised version of the weight for the boosting loss (Table 3), which corresponds to $p = 1$.

The above weight function view has at least three benefits. First, given the equivalence of the p-classification and p-norm push risk, we have some insight as to how the latter encourages solutions to maximise accuracy at the head of the ranked list: as p increases, the loss is seen to place relatively more weight on larger values of c. That is, we pay attention to those instances with high η values, as accurate modelling of these is essential for determining the behaviour about the boundary $\eta(x) = c$.

Second, we can design normalised versions of the p-classification loss that have more interpretable behaviour when $p \to \infty$. Evidently, $w_{\text{pcl}}(c; p)$ above tends to the trivial zero weight as $p \to \infty$, owing to the scaling factor of $(p+1)^{-1}$. Removing this scaling factor ensures that the weights are normalised for every $p > 0$, in the sense that $w(1/2; p) = 8$. Further, the resulting proper loss is easily verified to be

$$(\forall u \in [0,1])\ \lambda(u; p) = \left(\left(1 + \frac{1}{p}\right) \cdot \left(\frac{u}{1-u} \right)^{1-\frac{1}{p+1}}, (p+1) \cdot \left(\left(\frac{1-u}{u} \right)^{\frac{1}{p+1}} - 1 \right) \right),$$

with non-trivial limiting case as $p \to \infty$ of

$$(\forall u \in [0,1])\, \lambda(u; +\infty) = \left(\frac{u}{1-u}, -\log\frac{u}{1-u}\right).$$

When composed with the sigmoid link $\sigma(\cdot)$, this normalised family of proper losses results in the proper composite family

$$(\forall v \in \mathbb{R})\, \ell(v; p) = \left(\left(1 + \frac{1}{p}\right) \cdot e^{\frac{p}{p+1} \cdot v}, (p+1) \cdot (e^{-\frac{v}{p+1}} - 1)\right), \tag{60}$$

with non-trivial limiting case as $p \to \infty$ of

$$(\forall v \in \mathbb{R})\, \ell(v; +\infty) = (e^v, -v).$$

Third, the weight function view suggests a scheme of *designing new losses* for ranking the best, by constructing appropriate weight functions emphasising large values of η. We now pursue this idea.

9.6.2 STRICT PROPERNESS AND q-RTB LOSSES

We now study the design of q-RTB losses based on the theory of proper composite losses. We begin with the simple observation that any strictly proper composite loss is a q-RTB loss for every $q \in [0,1]$, by virtue of choosing the corresponding link function Ψ in Equation 52. More generally, the set of q-RTB losses is exactly the set of proper composite losses ℓ for which $\lambda = \ell \circ \Psi^{-1}$ is strictly proper on the interval $\eta \in [q,1]$ and (not necessarily strictly) proper on the interval $[0,q)$. This suggests a simple recipe for designing q-RTB losses; however, as we shall see, not enforcing strict properness poses computational challenges.

A key difficulty in designing q-RTB losses is the following. Suppose λ is a proper loss that is not strictly proper on an interval $I \subseteq [0,1]$. Then, λ is non-convex, and more importantly, cannot be made convex via a link function. To see this, recall that the canonical link function Ψ for a proper loss λ is the function for which $\lambda \circ \Psi^{-1}$ has the largest modulus of convexity. The weight function w of λ is related to the canonical link function Ψ by $w = \Psi'$. As λ is not strictly proper on I, we must have that $w \equiv 0$ on I. But then Ψ must be constant on I, and hence not invertible. Thus, to maintain convexity, it is essential to maintain strictness of the proper composite loss.

As a simple example, suppose λ is some strictly proper loss with weight function w. Now for some $q > 0$, consider the loss $\lambda_{\text{RTB}}(\cdot; q)$ with weight function

$$w_{\text{RTB}}(c; q) \doteq [\![c \geq q]\!] \cdot w(c). \tag{61}$$

This loss is not strictly proper on the interval $[0,q)$, and is strict on $[q,1]$. Therefore, the loss aims to accurately model instances x for which $\eta(x) \geq q$. We can explicitly compute the partial losses for $\lambda_{\text{RTB}}(\cdot; q)$ as follows.

Lemma 59 *Pick any proper loss λ with weight function w. For any $q > 0$, let $w_{\text{RTB}}(c; q)$ be the weight function given by Equation 61. Then, this weight has corresponding proper loss*

$$(\forall u \in [0,1])\, \lambda_{\text{RTB}}(-1, u; q) = \lambda_{-1}(u) - \lambda_{-1}(u \wedge q)$$
$$\lambda_{\text{RTB}}(+1, u; q) = \lambda_1(u \vee q).$$

Proof By Shuford's integral representation (Equation 10),

$$\begin{aligned}\lambda_{\text{RTB}}(-1, u; q) &= \int_0^1 [\![c < u]\!] \cdot [\![c \geq q]\!] \cdot c \cdot w(c)\, dc \\ &= \int_0^1 [\![c < u]\!] \cdot c \cdot w(c)\, dc - \int_0^1 [\![c < u]\!] \cdot [\![c < q]\!] \cdot c \cdot w(c)\, dc\end{aligned}$$

$$
\begin{aligned}
&= \int_0^1 [\![c < u]\!] \cdot c \cdot w(c)\, dc - \int_0^1 [\![c < u \wedge q]\!] \cdot c \cdot w(c)\, dc \\
&= \lambda_{-1}(u) - \lambda_{-1}(u \wedge q). \\
\lambda_{\mathrm{RTB}(q)}(+1, u; q) &= \int_0^1 [\![c > u]\!] \cdot [\![c \geq q]\!] \cdot (1 - c) \cdot w(c)\, dc \\
&= \int_0^1 [\![c > u \vee q]\!] \cdot (1 - c) \cdot w(c)\, dc \\
&= \lambda_1(u \vee q).
\end{aligned}
$$

■

When the prediction $u \geq q$, the partial losses of $\lambda_{\mathrm{RTB}}(\cdot; q)$ are unchanged from those of λ, barring a translation for λ_{-1}. However, when $u < q$, the partial loss for the positive class plateaus, whereas the partial loss for the negative class drops to zero. The resulting loss is clearly non-convex, and further, *no* invertible link function can be applied to make it convex: composing $\lambda_{\mathrm{RTB}}(\cdot; q)$ with a invertible link function Ψ yields a loss ℓ_{RTB} with partial losses

$$
\begin{aligned}
\ell_{\mathrm{RTB}}(-1, v; q) &= \ell_{-1}(v) - \ell_{-1}(v \wedge \Psi(q)) \\
\ell_{\mathrm{RTB}}(+1, v; q) &= \ell_1(v \vee \Psi(q)),
\end{aligned}
$$

which are not convex for any choice of Ψ.

A natural alternative to a loss that is not strictly proper is one that is "nearly" so, i.e. one whose weight function w is close to, but never exactly 0. However, this must be done with the following fact in mind: for any $\alpha > 0$, the proper loss with scaled weight function $\alpha \cdot w$ is simply the scaled loss $\alpha \cdot \lambda$. Thus, uniformly scaling a weight function does not affect the strict properness of the underlying loss. Scaling a loss on an interval $I \subset [0, 1]$ will however induce a qualitatively different loss: minimally, the new loss will be asymmetric, and have a non-trivially different set of Bayes-optimal scorers compared to the original loss. We now explore how proper composite losses can be designed to approximate a q-RTB loss.

9.6.3 PROPER COMPOSITE q-RTB SURROGATES

Our basic recipe for generating a q-RTB loss will be to combine the weight functions for two existing losses. Specifically, let w_ν, w_μ be weight functions corresponding to proper losses ν, μ. We assume that w_μ grows faster near 1 than w_ν does near 0, i.e. $\lim_{c \to 1} \frac{w_\mu(c)}{w_\nu(1-c)} > 1$. (We will typically be interested in the case where the limit is $+\infty$.) We now consider a hybrid weight function of the form

$$
\begin{aligned}
(\forall c \in (0, 1))\, \bar{w}(c; q) &\doteq \begin{cases} w_\nu(c) & \text{if } c < q \\ \alpha(q) \cdot w_\mu(c) & \text{if } c \geq q \end{cases} \\
&= w_\nu(c) \cdot [\![c < q]\!] + \alpha(q) \cdot w_\mu(c) \cdot [\![c \geq q]\!], \qquad (62)
\end{aligned}
$$

where $\alpha(q) = \frac{w_\nu(q)}{w_\mu(q)}$ so that the there is no discontinuity at $c = q$.

Since $\bar{w}(\cdot; q)$ is the sum of two weights, we can compute the corresponding proper losses for each component to get the form of the corresponding proper loss.

Lemma 60 *Pick any weight functions $w_\nu, w_\mu : [0, 1] \to \mathbb{R}_+$ with corresponding proper losses ν, μ. For any $q > 0$, let $\bar{w}(\cdot; q)$ be as per Equation 62. Then, the proper loss corresponding to this weight is*

$$
\begin{aligned}
(\forall u \in [0, 1])\, \bar{\lambda}_{-1}(u; q) &= \alpha(q) \cdot \mu_{-1}(u) + \nu_{-1}(u \wedge q) - \alpha(q) \cdot \mu_{-1}(u \wedge q) \\
\bar{\lambda}_1(u; q) &= \nu_1(u) + \alpha(q) \cdot \mu_1(u \vee q) - \nu_1(u \vee q),
\end{aligned} \qquad (63)
$$

where $\alpha(q) = \frac{w_\nu(q)}{w_\mu(q)}$.

Proof By Lemma 59, the weight $w_\nu(c) \cdot [\![c < q]\!] = w_\nu(c) - w_\nu(c) \cdot [\![c \geq q]\!]$ corresponds to the proper loss

$$\bar{\nu}(u) = (\nu_{-1}(u \wedge q), \nu_1(u) - \nu_1(u \vee q))$$

while the weight $\alpha(q) \cdot w_\mu(c) \cdot [\![c \geq q]\!]$ corresponds to the proper loss

$$\bar{\mu}(u) = (\alpha(q) \cdot \mu_{-1}(u) - \alpha(q) \cdot \mu_{-1}(u \wedge q), \alpha(q) \cdot \mu_1(u \vee q)).$$

Thus, the weight w corresponds to the sum of these losses, which is of the given form. ■

We may similarly combine proper composite losses corresponding to the underlying weights into a continuous proper composite loss corresponding to the weight $\bar{w}(\cdot; q)$.

Lemma 61 *Suppose that proper losses* ν, μ *have corresponding proper composite losses* ρ, κ *using invertible link functions* Ψ, Φ. *For any* $q > 0$, *let* $\bar{w}(\cdot; q)$ *be as per Equation 62. Then,* $\bar{w}(\cdot; q)$ *has corresponding proper composite loss* $\bar{\ell}$ *with components*

$$(\forall v \in \mathbb{R})\, \bar{\ell}_{-1}(v; q) = \begin{cases} \rho_{-1}(v) & \text{if } v < v_0(q) \\ \bar{\kappa}_{-1}(v; q) + \rho_{-1}(v_0) - \bar{\kappa}_{-1}(v_0; q) & \text{else} \end{cases}$$

$$\bar{\ell}_1(v; q) = \begin{cases} \rho_1(v) + \bar{\kappa}_1(v_0; q) - \rho_1(v_0) & \text{if } v < v_0(q) \\ \bar{\kappa}_1(v; q) & \text{else,} \end{cases}$$

where $v_0(q) = \Psi(q)$, *and*

$$\bar{\kappa}(y, v; q) = \alpha(q) \cdot \kappa\left(y, \frac{v - \beta(q)}{\gamma(q)}\right)$$
$$\beta(q) = \Psi(q) - \gamma(q) \cdot \Phi(q)$$
$$\gamma(q) = \frac{\Psi'(q)}{\Phi'(q)}.$$

Proof By definition, we have

$$\rho(y, v) = \nu(y, \Psi^{-1}(v))$$
$$\kappa(y, v) = \mu(y, \Phi^{-1}(v)).$$

Given any $q > 0$, we will construct a proper composite loss using the proper loss $\bar{\lambda}(\cdot; q)$ of Equation 63, composed with a link function Π that is a suitable combination of Ψ and Φ. The technical detail to attend to is to ensure there are no discontinuities with the resulting loss.

First, to ensure that the link for the two pieces of $\bar{w}$ coincide, with the same derivative at the threshold q, we modify the link for the second loss to

$$\bar{\Phi}(c; q) \doteq \gamma(q) \cdot \Phi(c) + \beta(q),$$

where

$$\beta(q) = \Psi(q) - \gamma(q) \cdot \Phi(q)$$
$$\gamma(q) = \frac{\Psi'(q)}{\Phi'(q)}.$$

Note that if $\Psi'(q) = \Phi'(q)$, this simplifies to the translated link $\Phi(c) + \Psi(q) - \Phi(q)$. This modified link has inverse

$$\bar{\Phi}^{-1}(v; q) = \Phi^{-1}\left(\frac{v - \beta(q)}{\gamma(q)}\right).$$

Now define the hybrid link

$$\Pi(c; q) \doteq [\![c < q]\!] \cdot \Psi(c) + [\![c \geq q]\!] \cdot \bar{\Phi}(c; q)$$

with corresponding inverse

$$\Pi^{-1}(v; q) = [\![v < v_0(q)]\!] \cdot \Psi^{-1}(v) + [\![v \geq v_0(q)]\!] \cdot \bar{\Phi}^{-1}(v; q)$$

for $v_0(q) = \Psi(q) = \bar{\Phi}(q; q)$. Then, the proper loss corresponding to $\bar{w}$ can be made proper composite, with

$$\begin{aligned}
\bar{\ell}(-1, v; q) &= \bar{\kappa}_{-1}(v; q) + \rho_{-1}(v \wedge v_0) - \bar{\kappa}_{-1}(v \wedge v_0; q) \\
\bar{\ell}(+1, v; q) &= \rho_1(v) + \bar{\kappa}_1(v \vee v_0; q) - \rho_1(v \vee v_0)),
\end{aligned}$$

where

$$\begin{aligned}
\bar{\kappa}(y, v; q) &= \alpha(q) \cdot \mu(y, (\bar{\Phi}^{-1}(v)) \\
&= \alpha(q) \cdot \kappa\left(y, \frac{v - \beta(q)}{\gamma(q)}\right).
\end{aligned}$$

■

The basic idea of the loss ℓ in Equation 63 is intuitive: one switches between choosing one of the underlying losses based on some threshold on the scores. The only additional ingredient is that scaling and translating of one of the losses to ensure continuity of the end result. We provide examples that illustrate the basic idea.

- Consider the weight function

$$w(c) = \begin{cases} \frac{1}{c \cdot (1-c)} & \text{if } c < q \\ \frac{2\sqrt{q \cdot (1-q)}}{c^{3/2} \cdot (1-c)^{3/2}} & \text{if } c \geq q, \end{cases}$$

  which is a hybrid of the weights for logistic and exponential loss. Using the sigmoid link yields the loss (for $q = \frac{1}{2}$)

$$\ell(v) = \left(\begin{cases} \log(1 + e^v) & \text{if } v < 0 \\ e^{v/2} + \log 2 - 1 & \text{if } v \geq 0 \end{cases}, \begin{cases} \log\left(1 + e^{-v}\right) - \log 2 + 1 & \text{if } v < 0 \\ e^{-v/2} & \text{if } v \geq 0 \end{cases} \right).$$

- Consider the weight function

$$w(c) = \begin{cases} 4 & \text{if } c < q \\ \frac{2\sqrt{q \cdot (1-q)}}{c^{3/2} \cdot (1-c)^{3/2}} & \text{if } c \geq q, \end{cases}$$

  which is a hybrid of the weights for square and exponential loss. Using an appropriate hybrid of the identity and sigmoid link yields the loss (for $q = \frac{1}{2}$)

$$\ell(v) = \left(\begin{cases} \frac{1}{2} \cdot (1 + v)^2 & \text{if } v < 0 \\ e^{v/2} - \frac{1}{2} & \text{if } v \geq 0 \end{cases}, \begin{cases} \frac{1}{2} \cdot (1 - v)^2 + \frac{1}{2} & \text{if } v < 0 \\ e^{-v/2} & \text{if } v \geq 0 \end{cases} \right).$$

- Consider for $p > 0$ the family of weights

$$w(c; p) = \frac{1}{c \cdot (1-c)^{2-\frac{1}{p+1}}},$$

which are similar to those employed by p-classification, except that the behaviour near $c = 0$ is fixed, and does not vary with p. Though not explicitly a hybrid, the weight is asymmetric, and thus the role of p is to tune the degree of focus on large values of η. For example, when $p = 1$, with the sigmoid link we have the proper composite loss

$$\ell(v) = \left(\frac{2}{\sqrt{\sigma(-v)}}, 2\tanh^{-1}(\sqrt{\sigma(-v)}) \right).$$

As another example, when $p = 3$, we get

$$\ell(v) = \left(\frac{4}{3\sigma(-v)^{3/4}}, 2\tanh^{-1}((\sigma(-v))^{1/4}) + 2\tan^{-1}((\sigma(-v))^{1/4}) \right).$$

It is clear that the above recipe can be applied for any suitable combination of weight functions, which we have argued to focus attention at the head of the ranked list. How do we choose amongst several such candidate weight functions? Put another way, can we characterise which hybrid weight function is the "best"? Answering such a question requires a precise sense in which one loss is "better" than another. This issue is only superficially simple, as even in binary classification for example, one cannot expect any given surrogate to be uniformly superior to all others in terms of resulting misclassification error (Reid and Williamson, 2010, Appendix A). Nonetheless, relating for example the weight function of a proper loss to generalisation ability in terms of a performance measure such as PTop would be of interest.

We emphasise also that the above represents just one recipe for generating suitable proper composite losses. If one can generate a suitable parametrised family of weights and link function generating a convex loss (e.g. the scaled p-classification loss of Equation 60), these would also be suitable for ranking the best problems.

9.7 Experiments with proper composite losses for ranking the best

We present experiments that assess the efficacy of several proper composite losses proposed in the previous section for the problem of maximising accuracy at the head of the ranked list. The aim of our experiments is *not* to position the new losses as a superior alternative to the existing p-classification and p-norm push approaches. Rather, we wish to demonstrate that the proper composite interpretation gives one way of generating a family of losses for this problem, with the p-classification loss being but one example of this family. An attraction of these losses is that they are simple to optimise using gradient-based methods, with complexity linear in the number of training examples (as opposed to methods that operate on pairs of examples).

To clarify the effect of the choice of loss and choice of risk, we consider all combinations of the three risk types considered in this paper – proper composite (Equation 14), bipartite (Equation 17), and p-norm push (Equation 56) – and the loss functions of interest. On the one hand, one expects the p-norm push risk to perform best when combined with a loss suitable for ranking the best. On the other hand, our analysis in the previous section indicates that there is promise in the minimisation of a suitable proper composite risk.

For our losses, we experiment with the standard logistic and exponential losses, as well as the p-classification loss. Based on our hybrid loss proposal in Lemma 61, we consider the following:

- The proper composite loss with weight $w(c) = \frac{1}{c \cdot (1-c)^{2-\frac{1}{p+1}}}$, and sigmoid link, which we term the "Log-p-classification Hybrid";
- The proper composite loss with weight being a hybrid of $\frac{1}{c \cdot (1-c)}$ and $\frac{1}{2 \cdot c^{3/2} \cdot (1-c)^{3/2}}$ about threshold $\frac{1}{p+1}$, and sigmoid link, which we term the "Log-Exp Hybrid";

- The proper composite loss with weight being a hybrid of 4 and $\frac{1}{2 \cdot c^{3/2} \cdot (1-c)^{3/2}}$ about threshold $\frac{1}{p+1}$, and link being a hybrid of the identity and sigmoid link, which we term the "Square-Exp Hybrid".

We compare these methods on four UCI datasets: `ionosphere`, `housing`, `german` and `car`. Each method was trained with a regularised linear model, where the training objective was minimised using L-BFGS (Nocedal and Wright, 2006, pg. 177). For each dataset, we created 5 random train-test splits in the ratio $2:1$. For each split, we performed 5-fold cross-validation on the training set to tune the strength of regularisation $\lambda \in \{10^{-6}, 10^{-5}, \ldots, 10^{2}\}$, and where appropriate the constant[23] $p \in \{1, 2, 4, 8, 16, 32, 64\}$. We then evaluated performance on the test set, and report the average across all splits. As performance measures, we used the AUC, ARR, DCG, AP, and PTop (Agarwal, 2011; Boyd et al., 2012). For all measures, a higher score is better. Parameter tuning was done based on the AP on the test folds.

The results are summarised in Tables 10 – 13, with the average ranks of each method with respect to each metric summarised in Table 14. No single method clearly outperforms all others in all metrics. However, we observe that the candidate proper composite losses are very competitive with the p-classification loss – the "Log-exp hybrid" and "Square-exp" hybrid in particular consistently perform comparably, and often better than p-classification. We especially find that the newly proposed proper composite losses perform well even when used as a surrogate loss as part of the bipartite risk. This confirms that the weight function perspective of the p-classification loss, and thus the p-norm push, is potentially practically useful for the design of losses suitable for ranking the best.

9.8 Existing work

Clémençon and Vayatis (2007) identified two subproblems in ranking the best instances. The first problem is determining which instances qualify as the best. The second problem is ranking amongst these identified best instances. The first problem can be thought of as simply recovering an appropriate level set of $\eta(x)$, without determining the specific $\eta(x)$ value, i.e. we simply wish to discover

$$\{x \in \mathcal{X} : \eta(x) \geq q\}.$$

When q is fixed, this can be solved by reducing the problem to cost-sensitive classification (Scott and Davenport, 2007). More generally, Clémençon and Vayatis (2007) considered the setting where q depends on the quantile of the scoring function. This poses challenges for analysis and estimation. The quantile version of the problem has been studied theoretically by (Clémençon and Vayatis, 2007), and (Boyd et al., 2012) gave a practical convex optimisation solution for the case of hinge loss. In both cases, the threshold q was specified as a quantile of the η.

The problem of ranking amongst the best instances with a quantile-based threshold was studied theoretically by Clémençon and Vayatis (2007), who proposed that the optimal univariate scoring function here must satisfy

$$s^*(x) \in \begin{cases} \{\eta(x)\} & \text{if } \eta(x) \geq q_\eta \\ [0, q_\eta) & \text{if } \eta(x) < q_\eta. \end{cases}$$

Observe that this is identical to our Equation 52, except that q is now a function of η. They showed that two "local" versions of the AUC criterion, one of which is related to the partial AUC mentioned earlier, are optimised by this scorer. To our knowledge, our analysis in terms of proper losses for the simpler case where q is a fixed constant has not been done before.

Ertekin and Rudin (2011, Theorem 1) showed that for the case of a linear hypothesis class and $p \geq 1$, the Bayes optimal scorer for the p-norm push coincides with the classification risk for the asymmetric p-classification loss function of Equation 59. The optimal scorer for this loss is easily checked to be $s^* = \frac{1}{p+1} \cdot (\sigma^{-1} \circ \eta)$, and so in the unrestricted hypothesis class setting, the result agrees with ours. Our result is

23. We have observed that the parameter p selected by cross-validation may not necessarily correspond to the one that gives best test set performance, possibly a result of the limited sizes of the datasets in consideration. Treating each choice of p as resulting in a separate loss might therefore reveal slightly different rankings of the (loss, risk) combinations we consider.

Method	**AUC**	**ARR**	**DCG**	**AP**	**PTop**
Proper Logistic	0.9113 ± 0.0208 (15)	0.0583 ± 0.0056 (10)	0.2192 ± 0.0050 (12)	0.9243 ± 0.0339 (15)	13.0000 ± 17.0880 (9)
Proper Exponential	0.9128 ± 0.0166 (14)	0.0585 ± 0.0056 (9)	0.2193 ± 0.0050 (11)	0.9262 ± 0.0318 (14)	12.8000 ± 12.9499 (10)
Proper P-Classification	0.9152 ± 0.0160 (9)	0.0598 ± 0.0053 (5)	0.2207 ± 0.0045 (6)	0.9349 ± 0.0232 (8)	11.6000 ± 8.8487 (12)
Proper Log-p-classification Hybrid	0.9034 ± 0.0220 (16)	0.0606 ± 0.0021 (3)	0.2208 ± 0.0020 (5)	0.9236 ± 0.0194 (16)	8.4000 ± 5.4129 (13)
Proper Log-Exp Hybrid	0.9240 ± 0.0180 (2)	0.0601 ± 0.0054 (4)	0.2211 ± 0.0046 (4)	0.9430 ± 0.0263 (2)	16.2000 ± 13.7004 (3)
Proper Square-Exp Hybrid	0.9153 ± 0.0110 (8)	0.0601 ± 0.0052 (4)	0.2211 ± 0.0041 (4)	0.9395 ± 0.0191 (4)	16.8000 ± 10.5688 (2)
Bipartite Logistic	0.9157 ± 0.0195 (6)	0.0587 ± 0.0057 (8)	0.2197 ± 0.0049 (10)	0.9316 ± 0.0315 (10)	14.8000 ± 15.1228 (5)
Bipartite Exponential	0.9149 ± 0.0149 (11)	0.0590 ± 0.0053 (7)	0.2198 ± 0.0046 (9)	0.9292 ± 0.0292 (13)	13.0000 ± 12.7475 (9)
Bipartite P-Classification	0.9151 ± 0.0287 (10)	0.0575 ± 0.0077 (11)	0.2188 ± 0.0070 (13)	0.9294 ± 0.0361 (12)	15.6000 ± 14.6731 (4)
Bipartite Log-p-classification Hybrid	0.9207 ± 0.0131 (3)	0.0612 ± 0.0027 (2)	0.2218 ± 0.0028 (2)	0.9407 ± 0.0172 (3)	16.8000 ± 14.2373 (2)
Bipartite Log-Exp Hybrid	0.9166 ± 0.0160 (5)	0.0596 ± 0.0055 (6)	0.2205 ± 0.0046 (7)	0.9341 ± 0.0277 (9)	14.6000 ± 13.1643 (6)
Bipartite Square-Exp Hybrid	**0.9284 ± 0.0273 (1)**	**0.0618 ± 0.0025 (1)**	**0.2227 ± 0.0025 (1)**	**0.9522 ± 0.0281 (1)**	**28.8000 ± 19.8293 (1)**
P-Norm Logistic	0.9129 ± 0.0182 (13)	0.0596 ± 0.0058 (6)	0.2204 ± 0.0050 (8)	0.9314 ± 0.0292 (11)	14.0000 ± 11.6833 (7)
P-Norm Exponential	0.9154 ± 0.0147 (7)	0.0598 ± 0.0053 (5)	0.2207 ± 0.0044 (6)	0.9354 ± 0.0222 (7)	12.0000 ± 9.5131 (11)
P-Norm P-Classification	0.9152 ± 0.0287 (9)	0.0575 ± 0.0077 (11)	0.2188 ± 0.0070 (13)	0.9294 ± 0.0362 (12)	15.6000 ± 14.6731 (4)
P-Norm Log-p-classification Hybrid	0.7893 ± 0.0618 (17)	0.0475 ± 0.0026 (12)	0.2018 ± 0.0031 (14)	0.8215 ± 0.0589 (17)	1.2000 ± 1.6432 (14)
P-Norm Log-Exp Hybrid	0.9167 ± 0.0167 (4)	0.0598 ± 0.0055 (5)	0.2207 ± 0.0047 (6)	0.9358 ± 0.0264 (6)	13.4000 ± 11.4586 (8)
P-Norm Square-Exp Hybrid	0.9144 ± 0.0122 (12)	0.0612 ± 0.0024 (2)	0.2217 ± 0.0023 (3)	0.9361 ± 0.0152 (5)	13.0000 ± 9.6177 (9)

Table 10: Results of various "ranking the best" methods on `ionosphere` dataset.

Method	**AUC**	**ARR**	**DCG**	**AP**	**PTop**
Proper Logistic	0.7597 ± 0.0415 (2)	0.0438 ± 0.0179 (10)	0.2068 ± 0.0209 (7)	0.1490 ± 0.0623 (8)	0.0000 ± 0.0000 (3)
Proper Exponential	0.7563 ± 0.0824 (3)	0.0625 ± 0.0580 (5)	0.2213 ± 0.0441 (2)	**0.1762 ± 0.0752 (1)**	**0.4000 ± 0.8944 (1)**
Proper P-Classification	0.7344 ± 0.0964 (10)	0.0364 ± 0.0125 (15)	0.1991 ± 0.0198 (15)	0.1404 ± 0.0628 (13)	0.0000 ± 0.0000 (3)
Proper Log-p-classification Hybrid	0.7254 ± 0.1002 (15)	0.0424 ± 0.0190 (11)	0.2037 ± 0.0243 (10)	0.1423 ± 0.0689 (11)	0.0000 ± 0.0000 (3)
Proper Log-Exp Hybrid	**0.7785 ± 0.0461 (1)**	0.0402 ± 0.0135 (12)	0.2045 ± 0.0172 (9)	0.1490 ± 0.0578 (8)	0.0000 ± 0.0000 (3)
Proper Square-Exp Hybrid	0.7498 ± 0.0729 (5)	0.0402 ± 0.0205 (12)	0.2021 ± 0.0241 (12)	0.1429 ± 0.0682 (10)	0.0000 ± 0.0000 (3)
Bipartite Logistic	0.7280 ± 0.1085 (13)	0.0616 ± 0.0589 (6)	0.2187 ± 0.0471 (5)	0.1707 ± 0.0839 (5)	**0.4000 ± 0.8944 (1)**
Bipartite Exponential	0.7306 ± 0.0882 (11)	**0.0652 ± 0.0578 (1)**	**0.2222 ± 0.0466 (1)**	0.1740 ± 0.0837 (3)	**0.4000 ± 0.8944 (1)**
Bipartite P-Classification	0.7282 ± 0.0889 (12)	0.0627 ± 0.0586 (4)	0.2198 ± 0.0471 (3)	0.1704 ± 0.0841 (6)	**0.4000 ± 0.8944 (1)**
Bipartite Log-p-classification Hybrid	0.7382 ± 0.0711 (7)	0.0585 ± 0.0512 (7)	0.2170 ± 0.0398 (6)	0.1645 ± 0.0666 (7)	0.2000 ± 0.4472 (2)
Bipartite Log-Exp Hybrid	0.7547 ± 0.0710 (4)	0.0636 ± 0.0578 (2)	**0.2222 ± 0.0445 (1)**	0.1760 ± 0.0760 (2)	**0.4000 ± 0.8944 (1)**
Bipartite Square-Exp Hybrid	0.7273 ± 0.1094 (14)	0.0632 ± 0.0590 (3)	0.2195 ± 0.0486 (4)	0.1709 ± 0.0874 (4)	**0.4000 ± 0.8944 (1)**
P-Norm Logistic	0.6987 ± 0.1159 (17)	0.0317 ± 0.0129 (17)	0.1913 ± 0.0213 (17)	0.1190 ± 0.0501 (17)	0.0000 ± 0.0000 (3)
P-Norm Exponential	0.7377 ± 0.0691 (8)	0.0440 ± 0.0185 (9)	0.2054 ± 0.0233 (8)	0.1442 ± 0.0621 (9)	0.0000 ± 0.0000 (3)
P-Norm P-Classification	0.7495 ± 0.0632 (6)	0.0368 ± 0.0132 (14)	0.1998 ± 0.0195 (13)	0.1414 ± 0.0609 (12)	0.0000 ± 0.0000 (3)
P-Norm Log-p-classification Hybrid	0.7354 ± 0.0834 (9)	0.0348 ± 0.0116 (16)	0.1970 ± 0.0183 (16)	0.1354 ± 0.0591 (15)	0.0000 ± 0.0000 (3)
P-Norm Log-Exp Hybrid	0.6875 ± 0.1557 (18)	0.0469 ± 0.0316 (8)	0.2023 ± 0.0206 (11)	0.1353 ± 0.0545 (16)	0.2000 ± 0.4472 (2)
P-Norm Square-Exp Hybrid	0.7055 ± 0.1290 (16)	0.0400 ± 0.0162 (13)	0.1996 ± 0.0215 (14)	0.1364 ± 0.0657 (14)	0.0000 ± 0.0000 (3)

Table 11: Results of various "ranking the best" methods on `housing` dataset.

Method	**AUC**	**ARR**	**DCG**	**AP**	**PTop**
Proper Logistic	0.8121 ± 0.0285 (7)	0.0393 ± 0.0040 (4)	0.1870 ± 0.0040 (5)	0.6236 ± 0.0637 (7)	2.4000 ± 1.9494 (4)
Proper Exponential	0.8131 ± 0.0311 (3)	0.0372 ± 0.0047 (13)	0.1855 ± 0.0040 (12)	0.6218 ± 0.0677 (10)	1.8000 ± 2.0494 (7)
Proper P-Classification	0.8115 ± 0.0282 (9)	0.0389 ± 0.0048 (7)	0.1867 ± 0.0038 (7)	0.6226 ± 0.0621 (9)	2.4000 ± 2.3022 (4)
Proper Log-p-classification Hybrid	0.8103 ± 0.0278 (13)	**0.0407 ± 0.0039 (1)**	**0.1883 ± 0.0036 (1)**	**0.6285 ± 0.0576 (1)**	**3.4000 ± 2.3022 (1)**
Proper Log-Exp Hybrid	0.8111 ± 0.0290 (10)	0.0379 ± 0.0040 (12)	0.1860 ± 0.0040 (11)	0.6205 ± 0.0666 (14)	2.0000 ± 2.0000 (6)
Proper Square-Exp Hybrid	0.8086 ± 0.0322 (16)	0.0391 ± 0.0037 (6)	0.1867 ± 0.0040 (7)	0.6188 ± 0.0661 (15)	2.2000 ± 1.7889 (5)
Bipartite Logistic	0.8136 ± 0.0299 (2)	0.0382 ± 0.0043 (11)	0.1863 ± 0.0043 (9)	0.6233 ± 0.0682 (8)	2.6000 ± 2.7019 (3)
Bipartite Exponential	0.8118 ± 0.0268 (8)	0.0393 ± 0.0037 (4)	0.1870 ± 0.0038 (5)	0.6216 ± 0.0631 (11)	2.4000 ± 1.9494 (4)
Bipartite P-Classification	0.8131 ± 0.0298 (3)	0.0393 ± 0.0037 (4)	0.1871 ± 0.0038 (4)	0.6245 ± 0.0657 (4)	2.2000 ± 1.6432 (5)
Bipartite Log-p-classification Hybrid	0.8101 ± 0.0296 (14)	0.0401 ± 0.0036 (3)	0.1878 ± 0.0036 (3)	0.6279 ± 0.0637 (2)	2.2000 ± 1.3038 (5)
Bipartite Log-Exp Hybrid	**0.8138 ± 0.0311 (1)**	0.0384 ± 0.0048 (10)	0.1864 ± 0.0039 (8)	0.6244 ± 0.0664 (5)	2.2000 ± 2.1679 (5)
Bipartite Square-Exp Hybrid	0.8127 ± 0.0304 (5)	0.0387 ± 0.0052 (8)	0.1867 ± 0.0042 (7)	0.6240 ± 0.0640 (6)	2.4000 ± 2.0736 (4)
P-Norm Logistic	0.8129 ± 0.0296 (4)	0.0392 ± 0.0049 (5)	0.1870 ± 0.0038 (5)	0.6240 ± 0.0639 (6)	2.8000 ± 2.5884 (2)
P-Norm Exponential	0.8105 ± 0.0277 (11)	0.0385 ± 0.0044 (9)	0.1863 ± 0.0035 (9)	0.6206 ± 0.0614 (13)	2.0000 ± 2.0000 (6)
P-Norm P-Classification	0.8095 ± 0.0275 (15)	0.0382 ± 0.0042 (11)	0.1861 ± 0.0033 (10)	0.6188 ± 0.0604 (15)	1.6000 ± 1.3416 (8)
P-Norm Log-p-classification Hybrid	0.8104 ± 0.0294 (12)	0.0404 ± 0.0032 (2)	0.1880 ± 0.0033 (2)	0.6270 ± 0.0645 (3)	2.2000 ± 1.3038 (5)
P-Norm Log-Exp Hybrid	0.8104 ± 0.0277 (12)	0.0384 ± 0.0044 (10)	0.1863 ± 0.0034 (9)	0.6209 ± 0.0614 (12)	2.0000 ± 2.0000 (6)
P-Norm Square-Exp Hybrid	0.8124 ± 0.0278 (6)	0.0389 ± 0.0047 (7)	0.1868 ± 0.0035 (6)	0.6244 ± 0.0602 (5)	2.2000 ± 1.9235 (5)

Table 12: Results of various "ranking the best" methods on `german` dataset.

Method	**AUC**	**ARR**	**DCG**	**AP**	**PTop**
Proper Logistic	0.9976 ± 0.0012 (2)	**0.1706 ± 0.0284 (1)**	**0.3411 ± 0.0275 (1)**	0.9391 ± 0.0370 (3)	**13.2000 ± 3.9623 (1)**
Proper Exponential	0.9976 ± 0.0012 (2)	0.1705 ± 0.0286 (2)	0.3410 ± 0.0277 (2)	0.9376 ± 0.0339 (5)	12.8000 ± 3.9623 (3)
Proper P-Classification	0.9968 ± 0.0022 (8)	0.1703 ± 0.0290 (4)	0.3405 ± 0.0283 (6)	0.9316 ± 0.0394 (13)	12.8000 ± 3.9623 (3)
Proper Log-p-classification Hybrid	0.9973 ± 0.0014 (5)	0.1705 ± 0.0288 (2)	0.3409 ± 0.0280 (3)	0.9356 ± 0.0355 (8)	13.0000 ± 3.9370 (2)
Proper Log-Exp Hybrid	0.9973 ± 0.0014 (5)	0.1696 ± 0.0282 (8)	0.3399 ± 0.0273 (10)	0.9274 ± 0.0455 (17)	11.8000 ± 4.3243 (6)
Proper Square-Exp Hybrid	0.9972 ± 0.0018 (6)	0.1691 ± 0.0277 (9)	0.3395 ± 0.0269 (11)	0.9265 ± 0.0597 (18)	11.8000 ± 5.7184 (6)
Bipartite Logistic	0.9976 ± 0.0013 (2)	0.1704 ± 0.0286 (3)	0.3409 ± 0.0276 (3)	0.9371 ± 0.0375 (6)	12.8000 ± 3.9623 (3)
Bipartite Exponential	0.9976 ± 0.0012 (2)	0.1704 ± 0.0287 (3)	0.3408 ± 0.0278 (4)	0.9364 ± 0.0348 (7)	12.2000 ± 4.1473 (5)
Bipartite P-Classification	**0.9977 ± 0.0012 (1)**	**0.1706 ± 0.0290 (1)**	**0.3411 ± 0.0281 (1)**	**0.9394 ± 0.0340 (1)**	12.4000 ± 4.1593 (4)
Bipartite Log-p-classification Hybrid	0.9975 ± 0.0014 (3)	0.1702 ± 0.0278 (5)	0.3406 ± 0.0268 (5)	0.9354 ± 0.0457 (10)	13.0000 ± 4.3589 (2)
Bipartite Log-Exp Hybrid	0.9975 ± 0.0012 (3)	0.1703 ± 0.0287 (4)	0.3408 ± 0.0277 (4)	0.9355 ± 0.0360 (9)	12.2000 ± 4.1473 (5)
Bipartite Square-Exp Hybrid	0.9973 ± 0.0017 (5)	0.1698 ± 0.0284 (7)	0.3401 ± 0.0275 (8)	0.9293 ± 0.0523 (15)	13.0000 ± 4.4159 (2)
P-Norm Logistic	0.9976 ± 0.0013 (2)	**0.1706 ± 0.0286 (1)**	**0.3411 ± 0.0277 (1)**	0.9392 ± 0.0384 (2)	**13.2000 ± 3.9623 (1)**
P-Norm Exponential	0.9968 ± 0.0021 (8)	0.1702 ± 0.0288 (5)	0.3405 ± 0.0282 (6)	0.9307 ± 0.0370 (14)	12.8000 ± 3.9623 (3)
P-Norm P-Classification	0.9968 ± 0.0020 (8)	0.1704 ± 0.0291 (3)	0.3406 ± 0.0285 (5)	0.9318 ± 0.0358 (12)	13.0000 ± 3.9370 (2)
P-Norm Log-p-classification Hybrid	0.9969 ± 0.0019 (7)	0.1698 ± 0.0281 (7)	0.3400 ± 0.0272 (9)	0.9278 ± 0.0459 (16)	12.4000 ± 4.3359 (4)
P-Norm Log-Exp Hybrid	0.9976 ± 0.0012 (2)	0.1705 ± 0.0284 (2)	0.3410 ± 0.0275 (2)	0.9388 ± 0.0351 (4)	12.8000 ± 3.9623 (3)
P-Norm Square-Exp Hybrid	0.9974 ± 0.0015 (4)	0.1700 ± 0.0280 (6)	0.3403 ± 0.0270 (7)	0.9320 ± 0.0498 (11)	**13.2000 ± 3.7683 (1)**

Table 13: Results of various "ranking the best" methods on `car` dataset.

Method	AUC	ARR	DCG	AP	PTop
Proper Logistic	6.5000	6.2500	6.2500	8.2500	4.2500
Proper Exponential	5.5000	7.2500	6.7500	7.5000	5.2500
Proper P-Classification	9.0000	7.7500	8.5000	10.7500	5.5000
Proper Log-p-classification Hybrid	12.2500	4.2500	4.7500	9.0000	4.7500
Proper Log-Exp Hybrid	4.5000	9.0000	8.5000	10.2500	4.5000
Proper Square-Exp Hybrid	8.7500	7.7500	8.5000	11.7500	4.0000
Bipartite Logistic	5.7500	7.0000	6.7500	7.2500	3.0000
Bipartite Exponential	8.0000	**3.7500**	4.7500	8.5000	4.7500
Bipartite P-Classification	6.5000	5.0000	5.2500	5.7500	3.5000
Bipartite Log-p-classification Hybrid	6.7500	4.2500	**4.0000**	**5.5000**	2.7500
Bipartite Log-Exp Hybrid	**3.2500**	5.5000	5.0000	6.2500	4.2500
Bipartite Square-Exp Hybrid	6.2500	4.7500	5.0000	6.5000	**2.0000**
P-Norm Logistic	9.0000	7.2500	7.7500	9.0000	3.2500
P-Norm Exponential	8.5000	7.0000	7.2500	10.7500	5.7500
P-Norm P-Classification	9.5000	9.7500	10.2500	12.7500	4.2500
P-Norm Log-p-classification Hybrid	11.2500	9.2500	10.2500	12.7500	6.5000
P-Norm Log-Exp Hybrid	9.0000	6.2500	7.0000	9.5000	4.7500
P-Norm Square-Exp Hybrid	9.5000	7.0000	7.5000	8.7500	4.5000

Table 14: Average ranks of various "ranking the best" methods for each performance measure across all datasets.

more general in the sense of being for an unrestricted hypothesis class, and uses proper loss techniques. The result of Ertekin and Rudin (2011) holds for the case of a linear (possibly misspecified) function class.

Variants of the p-norm push have been proposed, although the focus has been on algorithmic issues (Rudin, 2009; Agarwal, 2011; Li et al., 2014).

Cossock and Zhang (2008) proposed to use an importance weighting approach to the problem of focussing on the head of the ranked list, and showed that the DCG of a ranking can be bounded by importance weighted squared error.

10. Exact compositional reductions between classification and ranking

We have introduced several seemingly distinct problems above, among them classification, class-probability estimation, pairwise ranking, and bipartite ranking. We now map out the relationships between these problems[24]. Our focus is on whether the Bayes-optimal solution for one of these problems can be transformed to give the optimal solution for another problem. Formally, for some pair of problems (A, B) – where we understand a "problem" to mean a specification of a distribution and loss, and hence the Bayes-optimal solutions – we would like to know if

$$(\forall s \in \mathcal{S}^{B,*})\,(\exists f : \mathbb{R} \to \mathbb{R})\; f \circ s \in \mathcal{S}^{A,*}$$

and vice-versa. When the above is true, we have an *exact compositional reduction* from A to B, in the sense of problem A providing an optimal solution for problem B via a transformation f[25]. According to our definition,

24. We will focus on the case of 0-1 loss for bipartite ranking, as this is the canonical performance measure in most studies. Recalling that the ℓ-bipartite ranking risk has equivalent Bayes-optimal solutions to class-probability estimation for certain strictly proper composite ℓ, the results derived here for class-probability estimation can be translated to the ℓ-bipartite ranking risk as well.

25. In practice, one must consider the impact misspecified hypothesis classes and finite samples have on transforming the solution of one problem to another. The recent work of Narasimhan and Agarwal (2013b) considers regret and generalisation bounds to this end.

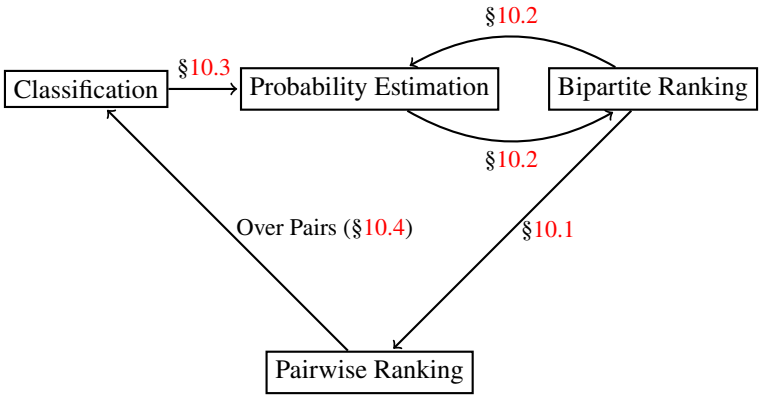


Figure 2: Relationships amongst various ranking and classification problems. An arrow $A \to B$ denotes that A is a special case of B, with the label on the arrow providing context on the relationship.

the transformation f may depend on the distribution specified by B. When this is so, the reduction is "weak" in the terminology of Narasimhan and Agarwal (2013b). When f is independent of B, the reduction is "strong" in the terminology of Narasimhan and Agarwal (2013b).

A sufficient condition for two problems to have the same Bayes-optimal solutions is the equivalence of the risks for the two problems. In some cases, this equivalence will be apparent from the specifications of the problem.

Figure 2 summarises the relationships amongst the various classification and ranking problems discussed in this paper. We now discuss these relationships in more detail.

10.1 Bipartite ranking $\subset$ Pairwise ranking

The bipartite ranking risk for a pair-scorer is defined with respect to a classification distribution D, while the pairwise ranking risk is defined with respect to a ranking distribution R. Therefore, in general for the two quantities to be equal, we need to have some relationship between the distributions $D \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ and $R \in \Delta_{\mathcal{X}\times\mathcal{X}\times\{\pm 1\}}$. The following shows that bipartite ranking is strictly a special case of pairwise ranking.

Proposition 62 *Pick any loss ℓ and scorer $s : \mathcal{X} \to \mathbb{R}$. Then, for every $D \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ there is some $R \in \Delta_{\mathcal{X}\times\mathcal{X}\times\{\pm 1\}}$ such that*

$$\mathbb{L}_{\mathrm{BR}}(s; D, \ell) = \mathbb{L}(\mathrm{Diff}(s); R, \ell).$$

Further, there exists some $R \in \Delta_{\mathcal{X}\times\mathcal{X}\times\{\pm 1\}}$ such that

$$(\nexists D \in \Delta_{\mathcal{X}\times\{\pm 1\}})\, \mathbb{L}_{\mathrm{BR}}(s; D, \ell) = \mathbb{L}(\mathrm{Diff}(s); R, \ell).$$

Proof The first statement is immediate by setting $R = D_{\mathrm{BR}}$, from Lemma 2.

For the second statement, suppose there were such a D. Then by Lemma 2, it must be true that

$$\mathbb{L}(\mathrm{Diff}(s); R, \ell) = \mathbb{L}(\mathrm{Diff}(s); D_{\mathrm{BR}}, \ell).$$

Thus, it must be true that $R = D_{\mathrm{BR}}$. However, if we choose R with base rate different from $\frac{1}{2}$, this cannot be possible, since D_{BR} has base rate $\frac{1}{2}$ by construction. We have a contradiction, and the result is shown. ■

Proposition 62 formalises the intuition that pairwise ranking is more general than bipartite ranking: there exist instances of the former (even when operating over decomposable pair-scorers) that cannot be solved by the latter.

10.2 Bipartite ranking = Class-probability estimation

We show that the Bayes-optimal solutions for bipartite ranking and class-probability estimation (with a strictly proper composite loss) may be transformed to one another. For bipartite ranking, the Bayes-optimal solution must be transformed in a distribution dependent manner (specifically, it must be calibrated with respect to the distribution).

Proposition 63 *Given any* $D = \langle M, \eta \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$ *and* $\ell \in \mathcal{L}_{\text{SPC}}$,

$$(\forall s \in \mathcal{S}^*(D, \ell))\, s \in \mathcal{S}^*_{\text{BR}}(D, \ell_{01})$$
$$(\forall s \in \mathcal{S}^*_{\text{BR}}(D, \ell_{01}))\, (\exists f^D) f^D \circ s \in \mathcal{S}^*(D, \ell).$$

Proof Pick $s \in \mathcal{S}^*(D, \ell)$. By Equation 44, this s is unique, and satisfies $s = \Psi \circ \eta$. Thus, by Corollary 43, $s \in \mathcal{S}^*_{\text{BR}}(D, \ell_{01})$.

Now pick $s \in \mathcal{S}^*_{\text{BR}}(D, \ell_{01})$. By Proposition 42, $\eta = \phi \circ s$ for some non-decreasing ϕ. Thus, by Lemma 74, the calibrated version $\text{Cal}(s; D)$ of s must equal η. Letting $f^D = \Psi \circ \text{Cal}(\cdot; D)$ gives the result. ■

The strict properness of the proper composite loss is essential for these results. Given a non-strictly proper composite loss, such as 0-1 loss, it is not true that *every* Bayes-optimal solution is also optimal for bipartite ranking, as we will now see.

10.3 Classification ⊂ Class-probability estimation

Class-probability estimation can be shown to be a more general problem than binary classification, in the sense that an optimal solution for the former can be transformed to one for the latter, but the other direction is only true for certain classes of distributions.

Proposition 64 *Given any* $D = \langle M, \eta \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$ *and* $\ell \in \mathcal{L}_{\text{SPC}}(\Psi)$,

$$(\forall s \in \mathcal{S}^*(D, \ell))\, (\exists f)\, f \circ s \in \mathcal{S}^*(D, \ell_{01})$$
$$(\forall s \in \mathcal{S}^*(D, \ell_{01}))\, (\exists f^D)\, f^D \circ s \in \mathcal{S}^*(D, \ell) \iff (\forall x \in \mathcal{X})\, \eta(x) \in \{a, b\},$$

where $a = b \neq \frac{1}{2}$ *or* $(2a - 1)(2b - 1) < 0$, *i.e.* η *is constant or takes on exactly two values on different sides of* $1/2$.

Proof For the first result, pick $s \in \mathcal{S}^*(D, \ell)$. By Equation 44, this s is unique, and satisfies

$$s = \Psi \circ \eta.$$

Since $2\eta - 1 \in \mathcal{S}^*(D, \ell_{01})$ (Equation 43), we can transform s to get an optimal solution for 0-1 loss.

For the second result, suppose that $\eta \in \{\frac{1}{2}, 1\}$. Then the scorer $[\![\eta \geq 1/2]\!] \equiv 1$ is in $\mathcal{S}^*(D, \ell_{01})$ by Equation 43. This scorer takes on exactly one value. Therefore, it cannot be transformed to a function that takes on two or more values. (Note that there may exist *some* scorers that can be transformed to give η, but this is not the proposition in question.) A similar argument shows that we cannot handle the case where η takes on three or more values.

Now suppose that η satisfies the given conditions, and pick any $s \in \mathcal{S}^*(D, \ell_{01})$. We know that $\text{sign}(s) = \text{sign}(2\eta - 1)$. Supposing without loss of generality that $a \leq b$, we have

$$\Psi \circ f^D \circ s = \Psi \circ \eta \in \mathcal{S}^*(D, \ell)$$

where $f^D(x) = b \cdot [\![x > 0]\!] + a \cdot [\![x < 0]\!]$. ■

The restriction on η above is satisfied by *separable* or *noiseless* distributions, where for every $x \in \mathcal{X}$, $\eta(x) \in \{0, 1\}$. The above thus shows the intuitive fact that in the absence of noise, classification and class-probability estimation are equivalent in terms of their end goals[26].

10.4 Classification over pairs = Pairwise ranking

We can confirm that pairwise ranking is equivalent to binary classification over the instance space $\mathcal{X} \times \mathcal{X}$ by simply comparing the risks for the two problems (Equations 19 and 14). This implies that bipartite ranking is also a special case of binary classification over $\mathcal{X} \times \mathcal{X}$, due to the relationship between bipartite and pairwise ranking established in §10.1.

10.5 Classification over singletons ⊂ Bipartite ranking

While bipartite ranking reduces to classification over *pairs*, it does not reduce to classification over *singletons*, except in special cases. We now present conditions for the equivalence of the Bayes-optimal solutions of the two problems.

Proposition 65 *Given any* $D = \langle M, \eta \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$,

$$(\forall s \in \mathcal{S}^*_{\mathrm{BR}}(D, \ell_{01}))\,(\exists f) f \circ s \in \mathcal{S}^*(D, \ell_{01})$$
$$(\forall s \in \mathcal{S}^*(D, \ell_{01}))\,(\exists f^D) f^D \circ s \in \mathcal{S}^*_{\mathrm{BR}}(D, \ell_{01}) \iff (\forall x \in \mathcal{X})\, \eta(x) \in \{a, b\},$$

where $a = b \neq \frac{1}{2}$ *or* $(2a - 1)(2b - 1) < 0$*, i.e.* η *is constant or takes on exactly two values on different sides of* $1/2$.

Proof The result follows by the established relationships between classification and class-probability estimation (§10.3), and class-probability estimation and bipartite ranking (§10.2). ■

As with class-probability estimation, the above shows that for separable distributions, bipartite ranking is equivalent to binary classification in terms of the end goal.

10.6 Relation to existing work

For the case of 0-1 loss, the fact that the bipartite ranking risk exactly equals a specific pairwise classification risk (and hence a specific pairwise ranking risk) is well known (Clémençon et al., 2008; Kotlowski et al., 2011; Uematsu and Lee, 2012; Agarwal, 2014). The derived ranking distribution D_{BR}, which explicitly specifies the pairwise ranking distribution for which this holds, has been invoked by Balcan et al. (2008); Kotlowski et al. (2011); Agarwal (2014), among others. Our generalisation to an arbitrary loss ℓ, while simple, appears novel.

Our study of the relationship between the bipartite and pairwise ranking problems differs from that of Balcan et al. (2008); Ailon and Mohri (2007) in at least two aspects. First, those works look at a *subset* version of bipartite ranking, where the goal is to rank a given subset of instances. Second, those works consider the goal of bipartite ranking to produce a univariate scorer rather than a pair-scorer. Therefore, they consider the question of how one can derive a univariate scoring function suitable for ranking from a classifier over pairs. The main result of Balcan et al. (2008) is that, given a classifier of pairs that achieves small classification risk, one can produce a univariate scorer with bipartite ranking risk that is worse by at most a factor of two.

11. Four bipartite ranking risks with equivalent minimisers

Consider the following approaches to producing a pair-scorer, given a strictly proper composite ℓ:

26. However, as noted earlier, for separable data the infimum in the definition of the Bayes risk may be unattainable for a strictly proper composite loss.

(1) $\mathrm{Diff}\left(\operatorname*{argmin}_{s\,:\,\mathcal{X}\to\mathbb{R}} \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim D}\left[e^{-\mathsf{Y}s(\mathsf{X})}\right]\right)$	(2) $\mathrm{Diff}\left(\operatorname*{argmin}_{s\,:\,\mathcal{X}\to\mathbb{R}} \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[e^{-(s(\mathsf{X})-s(\mathsf{X}'))}\right]\right)$
(3) $\operatorname*{argmin}_{s_{\mathrm{Pair}}\,:\,\mathcal{X}\times\mathcal{X}\to\mathbb{R}} \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[e^{-s_{\mathrm{Pair}}(\mathsf{X},\mathsf{X}')}\right]$	(4) $\mathrm{Diff}\left(\operatorname*{argmin}_{s\,:\,\mathcal{X}\to\mathbb{R}} \mathbb{E}_{\mathsf{X}'\sim Q}\left[\left(\mathbb{E}_{\mathsf{X}\sim P}\left[e^{-(s(\mathsf{X})-s(\mathsf{X}'))}\right]\right)^{p}\right]\right)$

Table 15: Four approaches for obtaining a pair-scorer in a bipartite ranking problem, using exponential loss. Our results show that the all approaches have the same theoretical minimiser.

(1) Minimise the ℓ-classification risk $\mathbb{L}(s; D, \ell)$, and $\mathrm{Diff}(\cdot)$ the result.

(2) Minimise the ℓ-bipartite ranking risk $\mathbb{L}_{\mathrm{BR}}(s; D, \ell)$ over all scorers, and $\mathrm{Diff}(\cdot)$ the result.

(3) Minimise the ℓ-pairwise ranking risk $\mathbb{L}(s_{\mathrm{Pair}}; D_{\mathrm{BR}}, \ell)$ over all pair-scorers.

(4) Minimise the p-norm push risk $\mathrm{Push}(s_{\mathrm{Pair}}; D, \ell_{\exp}, g^p)$ over *decomposable* pair-scorers.

Superficially, these appear very different: method (4) is the only one that departs from the standard conditional risk framework, method (3) is the only one to use a pair-scorer during minimisation, and method (1) is the only one to operate on single instances rather than pairs. It is thus surprising that our results provide conditions under which all methods have the *same* output; it is further surprising that the condition involves the choice of link function in the loss ℓ, which is typically chosen for computational rather than statistical reasons (Reid and Williamson, 2010).

Proposition 66 *Given any $D \in \Delta_{\mathcal{X}\times\{\pm1\}}$ and $\ell \in \mathcal{L}_{\mathrm{SPC}}(\Psi)$ with $\Psi \in \Sigma_{\mathrm{sig}}$, methods (1), (2) and (3) produce the same pair-scorer; if $\mathcal{X}$ is finite and $p = a - 1$ for $a > 1$, method (4) also produces the same pair-scorer.*

Proof By Equation 44 and Corollary 45, methods (1) and (2) produce the same scorer $\Psi\circ\eta$, up to a translation which is nullified by the Diff operator. By Equation 50, this pair-scorer is equivalent to that produced by method (3). Further, if $p = a - 1$ for $a > 1$, then by Proposition 57, method (4) returns $\Psi\circ\eta$ up to a translation which is nullified by the Diff operator. ■

In hindsight, these equivalences are not surprising by virtue of the Bayes-optimal scorer for each type of risk depending on the observation-conditional distribution η. They are not however *a priori* obvious, given how ostensibly different the risks appear. To illustrate these superficial differences, Table 15 provides a concrete example of the four methods when $\ell = \ell^{\exp}$ is the exponential loss, whose link $\Psi = \frac{1}{2}\sigma^{-1}$ satisfies the required condition of Proposition 66.

11.1 Implications of equivalences

The above shows the "equivalence" between four seemingly disparate risks, where our definition of "equivalent" is that two methods have the same optimal scorer. This does not imply that the methods are interchangeable in practice. A statistical caveat to these equivalences is that they ignore the issues of finite samples and a restricted function class. When one or both of these situations hold, it may be that one of these methods is more preferable. A computational caveat is that methods (2) – (4) rely on minimisation over pairs of examples. On a finite training set, this requires roughly quadratic complexity, compared to the linear complexity of method (1). These practical issues deserve investigation, but are beyond the scope of this paper.

This caveat in mind, we believe the results at least illuminate similarities between seemingly disparate approaches. For the problem of minimising the ℓ-bipartite risk for an appropriate surrogate ℓ, the above provides evidence that minimising the ℓ-classification risk is a suitable proxy. That is, performing class-probability estimation is a suitable proxy for ranking; this can be formalised with surrogate regret bounds (Agarwal, 2014; Narasimhan and Agarwal, 2013b). Similarly, for the problem of minimising the p-norm push objective, we have evidence that minimising the ℓ-classification or bipartite risk is a suitable proxy. As seen in §9.7, certain proper composite losses do indeed give comparable performance to the p-norm push.

Type	**Property** $(\forall x, x', x'' \in \mathcal{X})$
Total	$\neg(R(x,x') = -1 \text{ and } R(x',x) = -1)$
Anti-symmetric	$\neg(R(x,x') = +1 \text{ and } R(x',x) = +1)$
Transitive	$R(x,x') = +1 \text{ and } R(x',x'') = +1 \implies R(x,x'') = +1$
Reflexive	$R(x,x) = +1$
Continuous	for every pair of convergent sequences $(x_n), (y_n)$, $(\forall n \in \mathbb{N})\, R(x_n, y_n) = +1 \implies R\left(\lim_{n\to\infty} x_n, \lim_{n\to\infty} y_n\right) = +1$
Preorder	Reflexive, Transitive
Total preorder	Total, Transitive
Partial order	Reflexive, Transitive, Anti-symmetric
Total order	Total, Transitive, Anti-symmetric

Table 16: Some common types of binary relation R, and their defining properties.

11.2 Relation to existing work

Subsets of the above equivalences have been observed earlier for special cases. For the specific case of exponential loss and a linear hypothesis class, the equivalence between methods (1) and (2) was made by Ertekin and Rudin (2011, Theorem 3), Gao and Zhou (2012, Lemma 4), and Gao and Zhou (2015, Theorem 7), while the equivalence between method (1) and (4) was shown by Ertekin and Rudin (2011, Theorem 1); here, method (1) represents AdaBoost, and method (2) RankBoost. For the special case of convex margin losses, the equivalence between methods (2) and (3) was shown by Uematsu and Lee (2012). Ertekin and Rudin (2011, Section 4.4) conjectured a lack of an equivalence between LogitBoost and logistic regression, based on empirical findings; this apparent contradiction with our results is because the latter focusses on the case of a linear hypothesis class, which is possibly misspecified, while our results are for an unrestricted hypothesis class, or equally for a correctly specified one.

12. A utility representation perspective of bipartite ranking

Our final topic of study is what the theory of *utility representations* tells us about the bipartite ranking problem. In particular, we look at how this theory provides insight into the particular form of the observation-conditional distribution η_{Pair} of D_{BR} (Equation 47), which we saw had non-obvious implications for the derivation of Bayes-optimal scorers (§7.4.2) and surrogate regret bounds (§8).

12.1 Binary relations

Given a set $\mathcal{X}$, a *binary relation* $\mathcal{R}$ on $\mathcal{X}$ is some subset of $\mathcal{X} \times \mathcal{X}$. There are two standard ways of referring to a relation. The first is the operator $\succeq_{\mathcal{R}}$, with the semantics $x \succeq_{\mathcal{R}} x' \iff (x, x') \in \mathcal{R}$. The second is the function $r : \mathcal{X} \times \mathcal{X} \to \{\pm 1\}$, with the semantics $r(x, x') = +1 \iff (x, x') \in \mathcal{R}$. We will use these two representations interchangeably. Table 16 summarises some standard properties of binary relations, and examples of specific types of binary relations.

A *probabilistic binary relation* (sometimes called a *reciprocal* or *ipsodual binary relation*) (Baets et al., 2005, pg. 419) on a set $\mathcal{X}$ is some function $p : \mathcal{X} \times \mathcal{X} \to [0, 1]$ satisfying $p(x, x') + p(x', x) = 1$ for every $x, x' \in \mathcal{X}$. The pair $(\mathcal{X}, p)$ is sometimes referred to as a *forced choice pair comparison system* (Roberts, 1984,

pg. 273). Every probabilistic binary relation has an induced binary relation[27] $\succeq_{\mathcal{P}}$, with

$$x \succeq_{\mathcal{P}} x' \iff p(x, x') \geq \frac{1}{2}.$$

Suppose $g : \left[\frac{1}{2}, 1\right] \times \left[\frac{1}{2}, 1\right] \to [0, 1]$ is some function that is commutative (i.e. $g(x, y) = g(y, x) \, \forall x, y$) and monotone increasing in both arguments. A probabilistic binary relation p is said to be *g-stochastically transitive* (Baets et al., 2005, pg. 419) if

$$(\forall x, x', x'' \in \mathcal{X}) \, x \succeq_{\mathcal{P}} x' \text{ and } x' \succeq_{\mathcal{P}} x'' \implies p(x, x'') \geq g(p(x, x'), p(x', x'')).$$

Special cases of the function g correspond to popular notions of stochastic transitivity: the case $g(x, y) = 1/2$ is known as *weak stochastic transitivity* (Roberts, 1984, pg. 283), $g(x, y) = x \wedge y$ as *moderate stochastic transitivity* (Roberts, 1984, pg. 284), and $g(x, y) = x \vee y$ as *strong stochastic transitivity* (Roberts, 1984, pg. 284). It is easy to check that weak stochastic transitivity of p corresponds to transitivity of the associated binary relation $\succeq_{\mathcal{P}}$, provided ties are broken in favour of the relation existing.

12.2 Utility representations for binary relations

Let $\succeq_{\mathcal{R}}$ be a binary relation on $\mathcal{X}$. We say that $\succeq_{\mathcal{R}}$ has a *utility representation* if there is some $s : \mathcal{X} \to \mathbb{R}$ such that

$$(\forall x, x' \in \mathcal{X}) \, x \succeq_{\mathcal{R}} x' \iff s(x) \geq s(x').$$

We say that a probabilistic binary relation p has a *generalised utility representation* if there is some function $H : \mathbb{R} \times \mathbb{R} \to [0, 1]$ and some function $s : \mathcal{X} \to \mathbb{R}$ such that

$$(\forall x, x' \in \mathcal{X}) \, p(x, x') = H(s(x), s(x')), \tag{64}$$

where H is increasing in its first argument, decreasing in its second argument. In (Świtalski, 2003), it is additionally assumed that H is concave-convex. In psychometrics, such a model is sometimes also referred to as a simple scalability assumption (Krantz, 1967).

Table 17 summarises various special cases of the generalised utility representation that have been studied; see (Roberts, 1984, pg. 273 – 280) for details. The utility representations are ordered as follows:

$$\text{Strict} \subsetneq \text{Fechnerian} \subseteq \text{Strong} \subsetneq \text{Weak}.$$

Note that the inclusions above are all strict, except for that of Fechnerian and Strong representations. In fact, p has a strong utility representation if and only if it has a *restricted* Fechnerian utility representation, where the restriction is that the strictly monotone increasing ϕ for the Fechnerian representation is only defined on the Minkowski self-difference $f(\mathcal{X}) - f(\mathcal{X})$ (Roberts, 1984, pg. 279).

12.3 Existence of utility representations

A key question is whether one can characterise when a given (probabilistic) binary relation possesses a specific type of utility representation. This lets us relate the ordering properties of the relation – such as whether it is symmetric, transitive, *et cetera* – with its mathematical representation as a function.

A classical result of Debreu (1954) (which generalises a result of Eilenberg (1941)) characterises when a (non-probabilistic) binary relation may be expressed via a real-valued utility function. Recall that a binary relation $\succeq_{\mathcal{R}}$ is a total preorder if it is total (and hence reflexive) and transitive. The theorem is as follows.

27. More generally, one may define a family of relations, where each member specifies a different scheme for how ties – corresponding to $p(x, x') = \frac{1}{2}$ – are broken.

Utility type	Definition	Characterisation
Weak	$H(a,b) \geq \frac{1}{2} \iff a \geq b$	Weak stochastic transitivity, contour sets closed
Strong	$H(a,b) \geq H(c,d) \iff a - b \geq c - d$	Quadruple condition (under stochastic continuity)
Fechnerian	$H(a,b) = \phi(a-b)$, ϕ monotone	Quadruple or bicancellative condition
Strict	$H(a,b) = \frac{a}{a+b}$	Product rule

Table 17: Summary of various types of utility representations for binary probabilistic relations, from most to least general. By "definition" we mean the conditions required of H in the general utility representation of Equation 64. By "characterisation", we mean necessary and sufficient conditions on a probabilistic binary relation for the representation to hold. See text for details.

Proposition 67 ((Eilenberg, 1941), (Debreu, 1954), (Debreu, 1964), (Bridges and Mehta, 1995, pg. 46)) *Let $\mathcal{X}$ be a topological space that is either (a) connected and separable, or (b) second countable. Let $\succeq_{\mathcal{R}}$ be a binary relation on $\mathcal{X}$. Then $\succeq_{\mathcal{R}}$ defines a continuous total preorder if and only if there is some $s : \mathcal{X} \to \mathbb{R}$ such that*

$$x \succeq_{\mathcal{R}} x' \iff s(x) \geq s(x').$$

The result characterises precisely the class of binary relations that may be represented via a utility function $s : \mathcal{X} \to \mathbb{R}$. While the "if" direction is straightforward, the "only if" direction is not: it implies that *any* continuous total preorder can be perfectly represented via the standard ordering relation $\geq$ on the reals, for an appropriate choice of utility function $s : \mathcal{X} \to \mathbb{R}$.

For probabilistic binary relations, we will focus on the case of a strict utility representation[28], for which

$$H(a,b) = \frac{a}{a+b} = \frac{1}{1+\frac{b}{a}} = \sigma(\sigma^{-1}(a') - \sigma^{-1}(b')),$$

where $a' = \frac{a}{a+1}, b' = \frac{b}{b+1}$. Thus, if a probabilistic binary relation p possesses this representation,

$$(\exists s : \mathcal{X} \to \mathbb{R})\,(\forall x, x' \in \mathcal{X})\; p(x,x') = \sigma(\sigma^{-1}(s(x)) - \sigma^{-1}(s(x'))). \tag{65}$$

We have the following characterisation of the existence of a strict utility representation.

Proposition 68 ((Luce and Suppes, 1965, Theorem 48, pg. 350)) *Suppose η_{Pair} is a binary probabilistic relation. Then, η_{Pair} has a strict utility representation (Equation 65) if and only if it satisfies the product rule,*

$$(\forall x, x', x'' \in \mathcal{X})\, \eta_{\text{Pair}}(x,x') \cdot \eta_{\text{Pair}}(x',x'') \cdot \eta_{\text{Pair}}(x'',x) = \eta_{\text{Pair}}(x,x'') \cdot \eta_{\text{Pair}}(x'',x') \cdot \eta_{\text{Pair}}(x',x), \tag{66}$$

which, when $\eta_{\text{Pair}} \neq \{0,1\}$, is equivalently

$$(\forall x, x', x'' \in \mathcal{X})\, \eta_{\text{Pair}}(x,x') = \sigma(\sigma^{-1}(\eta_{\text{Pair}}(x,x'')) + \sigma^{-1}(\eta_{\text{Pair}}(x'',x'))).$$

The product rule encodes that the probability of an intransitive cycle of relations $\{x \succeq_{\mathcal{P}} x', x' \succeq_{\mathcal{P}} x'', x'' \succeq_{\mathcal{P}} x\}$ equals the probability of the cycle $\{x \succeq_{\mathcal{P}} x', x' \succeq_{\mathcal{P}} x'', x'' \succeq_{\mathcal{P}} x\}$. The product rule necessarily implies the quadruple condition introduced earlier (and hence strong stochastic transitivity), since a strict utility implies a Fechnerian one.

28. In psychometrics, the strict utility representation is referred to as the Bradley-Terry-Luce model (Luce, 1959; Bradley and Terry, 1952).

12.4 Implications for bipartite ranking

Bipartite ranking is fundamentally concerned with learning a scorer $s : \mathcal{X} \to \mathbb{R}$, and using this to rank instances. We can equivalently think of the problem as a special case of pairwise ranking, where we restrict attention to decomposable pair-scorers (Lemma 2). The pairwise ranking problem can in turn be interpreted as one of learning a binary relation: if $r : \mathcal{X} \times \mathcal{X} \to \{\pm 1\}$ is a binary relation on $\mathcal{X} \times \mathcal{X}$, then for any $R \in \Delta_{\mathcal{X} \times \mathcal{X} \times \{\pm 1\}}$ such that the Bayes-optimal 0-1 pair-scorers match r in sign, i.e.

$$(\forall x, x' \in \mathcal{X})\, r(x, x') = \operatorname{sign}(2\eta_{\text{Pair}}(x, x') - 1),$$

learning a pair-scorer from R is equivalent to learning the binary relation r. Thus, one can then ask what implications the decomposability restriction has on learning a binary relation. Observe that when thresholded at 0, the pair-scorer $\text{Diff}(s)$ yields a binary relation r over $\mathcal{X} \times \mathcal{X}$, with

$$r(x, x') = +1 \iff (\text{Diff}(s))(x, x') \geq 0 \iff s(x) \geq s(x').$$

By Proposition 67, the resulting relation r is a continuous, total preorder, as can be easily checked by the properties of $\text{Diff}(s)$. Less obviously, Proposition 67 implies that for *any* continuous, total preorder $\succeq_{\mathcal{R}}$ over $\mathcal{X} \times \mathcal{X}$ (for $\mathcal{X}$ with suitable topological properties), the problem of learning $\succeq_{\mathcal{R}}$ can be expressed as a bipartite ranking problem.

Lemma 69 *Let $\mathcal{X}$ be a topological space that is either (a) connected and separable, or (b) second countable. Let $\succeq_{\mathcal{R}}$ be a continuous, total preorder on $\mathcal{X}$. Then, there is a $D \in \Delta_{\mathcal{X} \times \{\pm 1\}}$ for which the Bayes-optimal bipartite scorer for 0-1 loss induces the same ranking as $\succeq_{\mathcal{R}}$.*

Proof For any continuous, total preorder $\succeq_{\mathcal{R}}$, there is a corresponding utility representation $s : \mathcal{X} \to \mathbb{R}$ (by Proposition 67). Pick any strictly monotone increasing $\phi : \mathbb{R} \to [0, 1]$, and let $\eta = \phi \circ s$. Further, pick any marginal distribution M over $\mathcal{X}$. Then, $D = \langle M, \eta \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$ with corresponding D_{BR} has

$$(\forall x, x' \in \mathcal{X})\ \operatorname{sign}(2\eta_{\text{Pair}}(x, x') - 1) = \operatorname{sign}(\eta(x) - \eta(x')) = \operatorname{sign}(s(x) - s(x')) = r(x, x'),$$

so that the Bayes-optimal bipartite scorer for 0-1 loss induces an ordering over $\mathcal{X} \times \mathcal{X}$ identical to $\succeq_{\mathcal{R}}$. ■

One can also consider the implications of utility representation theory for learning a *probabilistic* binary relation. As above, it is clear that the problem of learning a probabilistic binary relation p that satisfies the product rule can be expressed as finding $\mathcal{S}^*_{\text{BR}}(D, \ell)$ for some suitable D and proper composite ℓ with decomposable Bayes-optimal scorer. Equation 47 implies that for any D with derived distribution D_{BR}, the corresponding η_{Pair} posses a strict utility representation. Thus, setting η_{Pair} to coincide with the utility representation of p gives a means of learning the relation p.

Conversely, we can gain some insight as to the form of observation-conditional distribution η_{Pair} of D_{BR} is of the special form given by Equation 47, which in turn explains why the sigmoidal family of links arises in §7.4.2.

Lemma 70 *Given any $D \in \Delta_{\mathcal{X} \times \{\pm 1\}}$, the resulting $D_{\text{BR}} = \langle M_{\text{pair}}, \eta_{\text{Pair}} \rangle \in \Delta_{\mathcal{X} \times \mathcal{X} \times \{\pm 1\}}$ satisfies the product rule, and has*

$$(\forall x, x' \in \mathcal{X})\, \eta_{\text{Pair}}(x, x') = \sigma(\sigma^{-1}(s(x)) - \sigma^{-1}(s(x')))$$

for some $s : \mathcal{X} \to \mathbb{R}$.

Proof For $D = \langle P, Q, \pi \rangle$, from the construction of $D_{\text{BR}} = \langle M_{\text{pair}}, \eta_{\text{Pair}} \rangle = \langle P_{\text{pair}}, Q_{\text{pair}}, \pi_{\text{pair}} \rangle$, it is immediate that the probabilistic relation η_{Pair} it represents satisfies the product rule – this is because Bayes' rule and the nature of P_{pair} implies that

$$(\forall x, x' \in \mathcal{X})\, \eta_{\text{Pair}}(x, x') = \frac{P_{\text{pair}}(x, x') \cdot \pi_{\text{pair}}}{M_{\text{pair}}(x, x')}$$

$$= \frac{p(x) \cdot q(x')}{2\mu_{\text{pair}}(x, x')},$$

so that the condition for the product rule (Equation 66) may be written

$$(\forall x, x' \in \mathcal{X})\ \frac{p(x) \cdot q(x')}{\mu_{\text{pair}}(x, x')} \cdot \frac{p(x') \cdot q(x'')}{\mu_{\text{pair}}(x', x'')} \cdot \frac{p(x'') \cdot q(x)}{\mu_{\text{pair}}(x'', x)} = \frac{p(x) \cdot q(x'')}{\mu_{\text{pair}}(x, x'')} \cdot \frac{p(x'') \cdot q(x')}{\mu_{\text{pair}}(x'', x')} \cdot \frac{p(x') \cdot q(x)}{\mu_{\text{pair}}(x', x)}.$$

The numerators are clearly identical, and the denominators can be shown to be identical by explicit multiplication of the form of μ_{pair} in Appendix B. Consequently, Proposition 68 thus implies that η_{Pair} must possess a strict utility representation, meaning it is of the form

$$(\forall x, x' \in \mathcal{X})\ \eta_{\text{Pair}}(x, x') = \sigma(\sigma^{-1}(s(x)) - \sigma^{-1}(s(x')))$$

for some $s : \mathcal{X} \to \mathbb{R}$. ■

Thus, we have an explanation for the specific form of Equation 47 – it is due to the probabilistic relation implicitly underlying the bipartite ranking problem satisfying the product rule, in conjunction with the utility representation theorem (Proposition 68) for all such relations.

13. Conclusion and future work

We have provided a systematic study of the bipartite ranking problem through its statistical risk. In particular:

- We described a fundamental connection between bipartite raking and classification over pairs (§4).
- We studied several properties of the ROC curve, including a to our knowledge novel result (Proposition 13) on how dominance in ROC space implies dominance with respect to *any* proper composite loss.
- We derived a number of integral representations of the AUC (§5.6), relating them to the integral representation for proper losses, (§5.6.2).
- We related the Bayes-optimal bipartite risk to an f-divergence between product measures for the class-conditional densities (§6.2), generalising a result for the case of 0-1 loss due to Torgersen (1991).
- We determined the set of Bayes-optimal scorers for bipartite ranking (§7.3, §7.4, §7.5), and thus surrogate regret bounds for the minimisation of pairwise surrogates (§8).
- We studied Bayes-optimal scorers for the p-norm push risk (§9.5), and explained the risk in terms of the weight function for proper losses (§9.6.3). We used this to derive several new loss functions (§9.6), which demonstrated favourable empirical performance compared to the p-norm push risk on a number of real datasets (§9.7).
- We mapped out the relationships between bipartite ranking and other learning problems, such as pairwise ranking, class-probability estimation, and classification (§10.1, §10.2), and the equivalence between several seemingly disparate risks for popular approaches in bipartite ranking (§11).
- We showed how theorems of utility representation describe the class of ranking problems over pairs that can be modelled by bipartite ranking (§12.4).

Our results built upon the rich framework of proper composite losses, which are central to the study of class-probability estimation. We hope our results illustrate the value of the proper loss machinery in studying bipartite ranking problems.

We outline several possible areas for future work.

- *Beyond the Bayes-risk*. We acknowledge that the study of the risk is only an initial step in the broader understanding of the bipartite ranking problem. For example, in the study of the Bayes-optimal scorers, we have assumed no restrictions of the set of allowed scorers and pair-scorers. In practice, we have access to only a finite number of samples, and typically use a restricted function class. Understanding the impact this has on the risk equivalences we have established is of interest. For example, can we characterise when one risk is more preferable from a statistical or computational point of view?

- *From the AUC to AUPRC*. One may hope to extend our analysis to other popular performance metrics for bipartite ranking, such as the area under the precision-recall curve (AUPRC). While we briefly touched upon the AUPRC in §9.4.2, we deferred detailed study due to difficulties in expressing it as a risk. Understanding the properties this broader class of performance measures is of interest.

- *Extension to instance ranking*. A natural extension of our results would be the study of the more general instance ranking problem, where the label space $\mathcal{Y}$ is not binary. With suitable assumptions on $\mathcal{Y}$, it is possible to leverage some analysis from bipartite ranking for such problems (Clémençon et al., 2013). It is possible that tools from multi-class probability estimation (Vernet et al., 2011) may also be a useful tool in the study of this problem.

- *Converting pair-scorers to a univariate scorer*. Given a pair-scorer, it may be desirable to construct a procedure that converts it to a univariate scorer. Balcan et al. (2008) devises such a procedure and provides guarantees on its performance for the standard AUC. It is of interest whether this can be generalised.

- *Applications*. We believe there is scope for our analysis to be extended to problems where both class-probability estimation and bipartite ranking are heavily employed. For example, a challenging learning scenario is where one has access to only positive and unlabelled examples. Recent work has shown the value of class-probability estimation (Elkan and Noto, 2008) and bipartite ranking methods (Sellamanickam et al., 2011) for this task. We hope that our analysis can offer directions for theoretical and algorithmic development for this and related problems.

More broadly, by focussing on the statistical risk and abstracting away finite sample and optimisation issues, we have aimed to perform a *problem-oriented* rather than *method-oriented* analysis, as per Platt (1962). We believe this gives deeper insight into the connections between problems, and disentangles computational and statistical concerns. For example, akin to the distributional analysis of the probing reduction (Langford and Zadrozny, 2005) in Reid and Williamson (2009), we have seen in §10 the broad theoretical connections between bipartite ranking, class-probability estimation and classification, such as the use of a calibration transform to convert a scorer that ranks instances optimally to a class-probability estimator. We hope our results demonstrate the value and encourage further pursuit of this style of analysis for other learning problems.

Acknowledgments

This work was conducted as part of NICTA, and was supported by the Australian Research Council (RCW) and NICTA (AKM and RCW). NICTA was funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program. Thanks to Cynthia Rudin and the COLT referees for their helpful comments on a preliminary version of this work, to the JMLR referees for their valuable suggestions, and to Brendan van Rooyen for suggesting a simple proof of Lemma 72.

Appendix A. Assorted lemmas

We collect some assorted lemmas that are employed in proofs, but are not directly related to bipartite ranking.

Lemma 71 *Let $s : \mathcal{X} \to \mathbb{R}$ be any scorer. If*

$$(\forall t : \mathcal{X} \to \mathbb{R}) \int_{\mathcal{X}} s(x) \cdot t(x)\, dx = 0,$$

then s is zero almost everywhere.

Proof This is in fact a special case of the fundamental lemma of the calculus of variations, which in turn is a special case of the du Bois-Reymond lemma (Troutman, 1996, pg. 99), (Giaquinta and Hildebrandt, 2004, pg. 16), both of which only requires the statement hold for all infinitely differentiable t. We show the contrapositive. Suppose $s \neq 0$ on a set of nonzero measure. Then $t = s^2 : \mathcal{X} \to \mathbb{R}_+$ is > 0 on this same set of nonzero measure. Thus,

$$\int_{\mathcal{X}} s(x) \cdot s(x)\, dx = \int_{\mathcal{X}} s(x)^2\, dx > 0,$$

where the inequality holds by Folland (1999, Proposition 2.16). Thus, the statement holds. ■

Lemma 72 *Let $f, g : \mathcal{X} \to \mathbb{R}$. Then,*

$$(\forall x, x' \in \mathcal{X})\, f(x) < f(x') \implies g(x) < g(x')$$

if and only if $f = \phi \circ g$ for some non-decreasing $\phi : \mathbb{R} \to \mathbb{R}$.

Proof ($\Longleftarrow$). This is easily verified by the definition of ϕ being non-decreasing.

($\Longrightarrow$). We will construct such a non-decreasing ϕ. For any $y \in \mathrm{Im}(g)$, let

$$\mathcal{I}(y) = \{x \in \mathcal{X} : g(x) = y\}$$

be the preimage of y under g. For any $y \in \mathbb{R}$, let

$$\phi(y) \doteq \min\{f(x) : x \in \mathcal{I}(y)\}.$$

We will check that $f = \phi \circ g$, and that ϕ is non-decreasing.

First, note that for any $x, x' \in \mathcal{I}(y)$, by definition $g(x) = g(x')$. By the contrapositive of the assumption,

$$g(x) \geq g(x') \implies f(x) \geq f(x')$$

and by swapping x, x',

$$g(x) \leq g(x') \implies f(x) \leq f(x')$$

so that

$$g(x) = g(x') \implies f(x) = f(x').$$

Thus for any $x, x' \in \mathcal{I}(y)$, $f(x) = f(x')$. Thus, for any $x \in \mathcal{I}(y)$,

$$\phi(y) = f(x).$$

Now, for any $x_0 \in \mathcal{X}$,

$$\begin{aligned} \phi(g(x_0)) &= \min\{f(x) : x \in \mathcal{I}(g(x_0))\} \\ &= f(x_0). \end{aligned}$$

Thus, $f = \phi \circ g$. To see that ϕ is non-decreasing, pick $y < y'$, and $x \in \mathcal{I}(y), x' \in \mathcal{I}(y')$. Then $y = g(x) < g(x') = y'$. Since $g(x) < g(x')$ implies $f(x) = \phi(y) < \phi(y') = f(x')$, we see that $y < y' \implies \phi(y) < \phi(y')$. ■

Lemma 73 *Let $f, g : \mathcal{X} \to \mathbb{R}$. Then,*

$$(\forall x, x' \in \mathcal{X})\ \operatorname{sign}(f(x) - f(x')) = \operatorname{sign}(g(x) - g(x'))$$

if and only if $f = \phi \circ g$ for some strictly monotone increasing $\phi : \mathbb{R} \to \mathbb{R}$.

Proof We can equivalently write the condition as

$$(\forall x, x' \in \mathcal{X})\ f(x) < f(x') \iff g(x) < g(x').$$

Thus, by Lemma 72, $f = \phi_1 \circ g$ for some monotone increasing ϕ_1, and $g = \phi_2 \circ f$ for some monotone increasing ϕ_2. Thus $f = \phi_1 \circ \phi_2 \circ f$, and so $\phi_1 = \phi_2^{-1}$. This implies that ϕ_1 and ϕ_2 are invertible, or equivalently, that they both correspond to strictly monotone increasing transforms. ■

Lemma 74 *Given any $D = \langle M, \eta \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$, and any $s : \mathcal{X} \to \mathbb{R}$ such that $\eta = \phi \circ s$ for some non-decreasing ϕ,*

$$(\forall x \in \mathcal{X})\ \operatorname{Cal}(x; D, s) = \eta(x).$$

In particular, the above equation holds for any s such that $s = \phi \circ \eta$ for some strictly monotone ϕ.

Proof Assuming that the distribution of scores is discrete, for every $x \in \mathcal{X}$,

$$\begin{aligned}
\operatorname{Cal}(x; D, s) &= \mathbb{P}[\mathsf{Y} = 1 | s(\mathsf{X}) = s(x)] \\
&= \frac{\mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y}) \sim D} [\![\mathsf{Y} = 1, s(\mathsf{X}) = s(x)]\!]}{\mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y}) \sim D} [\![s(\mathsf{X}) = s(x)]\!]} \\
&= \frac{\mathop{\mathbb{E}}_{\mathsf{X} \sim M} \left[\eta(\mathsf{X}) \cdot [\![s(\mathsf{X}) = s(x)]\!]\right]}{\mathop{\mathbb{E}}_{\mathsf{X} \sim M} \left[[\![s(\mathsf{X}) = s(x)]\!]\right]} \\
&= \frac{\mathop{\mathbb{E}}_{\mathsf{X} \sim M} \left[\phi(s(\mathsf{X})) \cdot [\![s(\mathsf{X}) = s(x)]\!]\right]}{\mathop{\mathbb{E}}_{\mathsf{X} \sim M} \left[[\![s(\mathsf{X}) = s(x)]\!]\right]} \text{ by assumption on } \eta \\
&= \frac{\mathop{\mathbb{E}}_{\mathsf{X} \sim M} \left[\phi(s(x)) \cdot [\![s(\mathsf{X}) = s(x)]\!]\right]}{\mathop{\mathbb{E}}_{\mathsf{X} \sim M} \left[[\![s(\mathsf{X}) = s(x)]\!]\right]} \\
&= \phi(s(x)) \cdot \frac{\mathop{\mathbb{E}}_{\mathsf{X} \sim M} \left[[\![s(\mathsf{X}) = s(x)]\!]\right]}{\mathop{\mathbb{E}}_{\mathsf{X} \sim M} \left[[\![s(\mathsf{X}) = s(x)]\!]\right]} \\
&= \phi(s(x)) \\
&= \eta(x).
\end{aligned}$$

When the distribution of scores is continuous, we repeat the above, but using Dirac delta instead of indicator functions. ■

Appendix B. Properties of the derived ranking distribution D_{BR}

Suppose we have a distribution $D = \langle P, Q, \pi \rangle = \langle M, \eta \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$. Assume that M, P, Q have densities μ, p, q. We summarise some properties of the resulting distribution over pairs, $D_{\text{BR}} \in \Delta_{\mathcal{X} \times \mathcal{X} \times \{\pm 1\}}$. We will

associate this distribution with the random variable triplet $(\mathsf{X}, \mathsf{X}', \mathsf{Z})$. By definition, we have

$$\begin{aligned}
\mathbb{P}[\mathsf{Z} = 1] &= \frac{1}{2} \\
p_{\mathsf{X}|\mathsf{Z}=z}(x) &= [\![z = 1]\!] \cdot p(x) + [\![z = -1]\!] \cdot q(x) \\
p_{\mathsf{X}'|\mathsf{Z}=z}(x') &= [\![z = 1]\!] \cdot q(x') + [\![z = -1]\!] \cdot p(x').
\end{aligned}$$

From these, we may derive other marginals and conditionals for D_{BR}, and relate them to those of D:

$$\begin{aligned}
p_{\mathsf{X},\mathsf{X}'|\mathsf{Z}=z}(x, x') &= p_{\mathsf{X}|\mathsf{Z}=z}(x) \cdot p_{\mathsf{X}'|\mathsf{Z}=z}(x') \\
&= [\![z = 1]\!] \cdot p(x) \cdot q(x') + [\![z = -1]\!] \cdot p(x') \cdot q(x) \\
p_{\mathsf{X},\mathsf{X}'}(x, x') &= \frac{p(x) \cdot q(x') + p(x') \cdot q(x)}{2} \\
&= \frac{1}{2\pi(1-\pi)} \cdot \mu(x) \cdot \mu(x') \cdot (\eta(x) \cdot (1 - \eta(x')) + \eta(x') \cdot (1 - \eta(x))) \\
p_{\mathsf{X}}(x) &= \frac{p(x) + q(x)}{2} \\
p_{\mathsf{X}|\mathsf{X}'=x'}(x) &= \frac{p(x) \cdot q(x') + p(x') \cdot q(x)}{p(x) + q(x)} \\
\mathbb{P}[\mathsf{Z} = 1|\mathsf{X} = x] &= \frac{p(x)}{p(x) + q(x)} \\
&= \sigma(\sigma^{-1}(\eta(x)) - \sigma^{-1}(\pi)) \\
\mathbb{P}[\mathsf{Z} = 1|\mathsf{X} = x, \mathsf{X}' = x'] &= \frac{p(x) \cdot q(x')}{p(x) \cdot q(x') + p(x') \cdot q(x)} \\
&= \frac{1}{1 + \frac{q(x)}{p(x)} \cdot \frac{p(x')}{q(x')}} \\
&= \sigma(\sigma^{-1}(\eta(x)) - \sigma^{-1}(\eta(x'))).
\end{aligned}$$

The last identity follows because

$$\sigma^{-1}(\eta(x)) = \log \frac{\pi}{1-\pi} + \log \frac{p(x)}{q(x)}.$$

Thus, the distributions of primary interest in the paper are:

$$\begin{aligned}
P_{\mathrm{pair}} &= P \times Q \\
Q_{\mathrm{pair}} &= Q \times P \\
\pi_{\mathrm{pair}} &= \frac{1}{2} \\
M_{\mathrm{pair}} &= \frac{P \times Q + Q \times P}{2} \\
(\forall x, x' \in \mathcal{X})\, \eta_{\mathrm{Pair}}(x, x') &= \sigma(\sigma^{-1}(\eta(x)) - \sigma^{-1}(\eta(x'))).
\end{aligned}$$

Appendix C. Properties of the ROC curve

For completeness we present two well known results about the derivative of the ROC curve, and its relationship to the optimal threshold for cost-sensitive learning.

Proof [Proof of Proposition 8] By the chain rule,

$$(\forall \alpha \in (0, 1))\, \rho'(\alpha) = (\mathrm{FPR}^{-1})'(\alpha) \cdot \mathrm{TPR}'(\mathrm{FPR}^{-1}(\alpha))$$

$$= \frac{\mathrm{TPR}'(\mathrm{FPR}^{-1}(\alpha))}{\mathrm{FPR}'(\mathrm{FPR}^{-1}(\alpha))}$$
$$= \frac{p_S(\mathrm{FPR}^{-1}(\alpha))}{q_S(\mathrm{FPR}^{-1}(\alpha))} \text{ by Equation 22.}$$

By Bayes' rule, for a random variable $\mathsf{S} \sim S$ for S the distribution of scores,

$$(\forall a \in \mathbb{R})\, p_S(a) = p_{\mathsf{S}|\mathsf{Y}=1}(a) = \frac{\mathbb{P}[\mathsf{Y}=1|\mathsf{S}=a] \cdot \mathbb{P}[\mathsf{S}=a]}{\mathbb{P}[\mathsf{Y}=1]},$$

and similarly for q_S. Thus

$$\frac{p_S(a)}{q_S(a)} = \frac{1-\pi}{\pi} \cdot \frac{\mathbb{P}[\mathsf{Y}=1|\mathsf{S}=a]}{1-\mathbb{P}[\mathsf{Y}=1|\mathsf{S}=a]},$$

and by definition, $\mathrm{Prb}(a; D, s) = \mathbb{P}[\mathsf{Y}=1|\mathsf{S}=a]$. ■

Proof [Proof of Proposition 11] Let

$$(\forall t \in \mathbb{R})\, R(t) \doteq \mathbb{L}(s-t; D, \ell_{\mathrm{CS}(c)})$$
$$= \pi \cdot (1-c) \cdot \mathrm{FNR}(t) + (1-\pi) \cdot c \cdot \mathrm{FPR}(t)$$

be the risk for a fixed scorer s when using a threshold t. Pick any optimal threshold $t_0 \in t^*(c; D, s)$. This must satisfy

$$0 = R'(t_0) = \pi \cdot (1-c) \cdot \mathrm{FNR}'(t_0) + (1-\pi) \cdot c \cdot \mathrm{FPR}'(t_0),$$

i.e.

$$\frac{\pi}{1-\pi} \cdot \frac{\mathrm{TPR}'(t_0)}{\mathrm{FPR}'(t_0)} = \frac{c}{1-c}.$$

This is exactly

$$\rho'(\mathrm{FPR}(t_0)) = \frac{c}{1-c} \cdot \frac{1-\pi}{\pi}.$$

By Proposition 8, this implies

$$\frac{\mathrm{Prb}(t_0)}{1-\mathrm{Prb}(t_0)} = \frac{c}{1-c}.$$

When $\mathrm{Prb}(\cdot\,; D, s)$ is invertible, we thus have

$$t_0 = \mathrm{Prb}^{-1}(c),$$

and since the choice of t_0 was arbitrary from the set of optimal thresholds, we conclude that $t^*(c) = \{(\mathrm{Prb})^{-1}(c)\}$. ■

Appendix D. Relationship between the ROC curve and the Neyman-Pearson problem

The maximal ROC curve is intimately related to the following classical hypothesis testing problem. Suppose we have (known) probability distributions P_{+1}, P_{-1} over an instance space $\mathcal{X}$, with densities p_{-1}, p_{+1} with respect to some reference measure (this could simply be $(P_{+1}+P_{-1})/2$). We are given a sample x drawn from one of $P_{\{\pm 1\}}$. We wish to determine whether or not $i = 1$, i.e. conduct a hypothesis test between $H_0 : i = -1$ and $H_1 : i = +1$. The Neyman-Pearson problem (Lehmann and Romano, 2005, pg. 59) asks for the test that has the most *power* in discriminating between the two alternatives, assuming the false positive rate is fixed at some value $\alpha \in [0, 1]$.

Definition 75 (Neyman-Pearson problem) *Pick any $\pi \in (0,1)$, and let $D = \langle P_{+1}, P_{-1}, \pi \rangle \in \Delta_{\mathcal{X}\times\{\pm 1\}}$. For a fixed $\alpha \in [0,1]$, the Neyman-Pearson problem is*

$$\max_{h \in \{\pm 1\}^{\mathcal{X}}} \mathrm{TPR}(h; D) \,:\, \mathrm{FPR}(h; D) \leq \alpha,$$

where in an abuse of notation

$$\mathrm{TPR}(h; D) = \mathbb{P}_{\mathsf{X} \sim P_{+1}}\left[h(\mathsf{X}) = +1\right]$$
$$\mathrm{FPR}(h; D) = \mathbb{P}_{\mathsf{X} \sim P_{-1}}\left[h(\mathsf{X}) = +1\right].$$

The optimal classifier $h^*(x)$ is called the uniformly most powerful test *at α.*

From a learning perspective, a test h is simply a classifier $h : \mathcal{X} \to \{\pm 1\}$ that specifies which of the two hypotheses is preferred. Further, we can view the densities as being class-conditionals $p_y(x) = p_{\mathsf{X}|\mathsf{Y}=y}(x)$. Thus, given a distribution $D = \langle P, Q, \pi \rangle \in \Delta_{\mathcal{X}\times\{\pm 1\}}$, the Neyman-Pearson problem arises when we wish to find a classifier that has maximal true positive rate for a fixed false positive rate.

D.1 The Neyman-Pearson lemma

The Neyman-Pearson lemma (Lehmann and Romano, 2005, pg. 60) specifies the optimal solution to the Neyman-Pearson problem.

Lemma 76 (Neyman-Pearson lemma) *Pick any $\pi \in (0,1)$, and let $D = \langle P_{+1}, P_{-1}, \pi \rangle \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ where $P_{\{\pm 1\}}$ have densities $p_{\{\pm 1\}}$ with respect to some reference measure. For any $\alpha \in [0,1]$, the uniformly most powerful test at α is*

$$h^*(x; \alpha, D) = \left[\!\!\left[\frac{p_{+1}(x)}{p_{-1}(x)} \geq t^*(\alpha; D) \right]\!\!\right]$$

where $t^*(\alpha; D)$ is such that the classifier achieves desired false positive rate,

$$\mathrm{FPR}(h^*(x; \alpha, D); D) = \alpha.$$

For completeness, we present a proof based on Lagrange multipliers, following Hippenstiel (2001, pg. 92).

Proof Given a classifier h, let

$$\mathcal{A}_h \doteq \{x \in \mathcal{X} \,:\, h(x) = +1\}.$$

Then,

$$\mathrm{TPR}(h; D) = \mathbb{P}_{\mathsf{X} \sim P_+}\left[\mathsf{X} \in \mathcal{A}_h\right]$$
$$\mathrm{FPR}(h; D) = \mathbb{P}_{\mathsf{X} \sim P_-}\left[\mathsf{X} \in \mathcal{A}_h\right].$$

Thus, the problem is equivalent to

$$\max_{\mathcal{A} \subseteq \mathcal{X}} \int_{x \in \mathcal{A}} p_+(x)\, dx \text{ subject to } \int_{x \in \mathcal{A}} p_-(x)\, dx \leq \alpha.$$

Now consider the Lagrangian

$$\mathcal{L}(\mathcal{A}, \lambda) = \int_{x \in \mathcal{A}} (p_+(x) - \lambda \cdot p_-(x))\, dx + \lambda \cdot \alpha.$$

Clearly, the $\mathcal{A}$ which maximises $\mathcal{L}$ is such that the integrand is always nonnegative:

$$\mathcal{A}^*(\lambda^*) = \left\{ x \in \mathcal{X} \,:\, \frac{p_+(x)}{p_-(x)} \geq \lambda^* \right\} = \left\{ x \in \mathcal{X} \,:\, \phi\left(\frac{p_+(x)}{p_-(x)}\right) \geq \phi(\lambda^*) \right\},$$

for any ϕ strictly monotone increasing. Thus, the optimal test or classifier is based on thresholding a scorer of the form

$$s^*(x) = \phi\left(\frac{p_{+1}(x)}{p_{-1}(x)}\right)$$

with the threshold λ^* being the solution to the equation

$$\alpha = \int_{x \in \mathcal{A}^*(\lambda^*)} p_-(x).$$

■

In practical learning settings, the optimal solution to the Neyman-Pearson problem cannot be computed as it assumes full knowledge of the underlying distributions. A natural approach is to use empirical versions of the appropriate distributions. For a fixed false positive rate α, various optimisation schemes have been proposed such as neural network based density estimation (Streit, 1990), SVMs (Davenport et al., 2010) and non-convex optimisation (Gasso et al., 2011).

D.2 Implications of the Neyman-Pearson lemma

From a hypothesis testing perspective, the optimal scoring function is seen to be a strictly monotone increasing transform of the *likelihood ratio*

$$(\forall x \in \mathcal{X})\, \Lambda(x) \doteq \frac{p_{+1}(x)}{p_{-1}(x)}.$$

From an ROC perspective, the Neyman-Pearson problem can be seen as picking a particular point on the horizontal axis (by virtue of fixing the FPR), and asking for the scoring function that yields the maximum value along the vertical axis (by virtue of maximising the TPR). The Neyman-Pearson lemma concludes that this is achieved by a strictly monotone increasing transformation of $\Lambda(x)$, regardless of the FPR value. Thus, maximising the AUC can be seen as solving a Neyman-Pearson problem for every possible false positive rate, as in Corollary 16.

Appendix E. Bayes-optimal classification risk and regret

A classical result establishes that the Bayes 0-1 risk may be expressed in terms of the variational divergence between the class-conditional distributions P and Q (Devroye et al., 1996, pg. 14). This gives an interpretation of the Bayes 0-1 risk, which is the inherent "difficulty" of the problem, in terms of amount of overlap between the distributions for the two classes. In fact, the variational divergence can be replaced by any f-divergence, with the Bayes 0-1 risk being replaced by the Bayes ℓ-classification risk for suitable ℓ.

Proposition 77 ((Österreicher and Vajda, 1993), (Reid and Williamson, 2011, Theorem 9)) *For any $D = \langle P, Q, \pi \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$, convex $f : \mathbb{R}_+ \to \mathbb{R}$, and loss ℓ with conditional Bayes risk*

$$(\forall \eta \in (0,1))\, L^*(\eta; \ell) = -\frac{1-\eta}{1-\pi} \cdot f\left(\frac{1-\pi}{\pi} \cdot \frac{\eta}{1-\eta}\right),$$

the Bayes-risk can be written

$$\mathbb{L}^*(D, \ell) = L^*(\pi; \ell) - \mathbb{I}_f(P, Q). \tag{67}$$

Conversely, Equation 67 holds for any $D = \langle P, Q, \pi \rangle \in \Delta_{\mathcal{X} \times \{\pm 1\}}$, loss ℓ with concave conditional Bayes risk $L^*(\cdot; \ell) : [0,1] \to \mathbb{R}_+$, and $f : \mathbb{R}_+ \to \mathbb{R}$ defined by

$$(\forall t \in \mathbb{R})\, f(t) = L^*(\pi; \ell) - (\pi \cdot t + 1 - \pi) \cdot L^*\left(\frac{\pi \cdot t}{\pi \cdot t + 1 - \pi}; \ell\right).$$

Recalling that for a proper composite loss ℓ with underlying proper loss λ, the conditional Bayes risks coincide i.e. $L^*_\ell = L^*_\lambda$, we see that for a proper loss λ Proposition 77 holds for *any* choice of proper composite ℓ resulting from the composition of λ with an invertible link function Ψ.

One can also relate the regret with respect to any proper composite loss to an appropriate generative Bregman divergence.

Proposition 78 ((Buja et al., 2005; Reid and Williamson, 2011)) *For any $D = \langle M, \eta\rangle \in \Delta_{\mathcal{X}\times\{\pm 1\}}$, $\ell \in \mathcal{L}_{\mathrm{SPC}}(\Psi)$, and scorer $s : \mathcal{X} \to \mathbb{R}$,*

$$\mathrm{regret}(s; D, \ell) = \mathbb{B}_{-L^*}(\eta, \Psi^{-1} \circ s)$$

where in an abuse of notation $L^* = L^*(\cdot; \ell)$.

Proposition 78 shows that if a scorer has low ℓ-risk with respect to some proper composite loss, then $\hat{\eta} = \Psi^{-1} \circ s$ is a good estimate of η in a precise sense: it has low average Bregman divergence to η.

Appendix F. Interpretation of Uematsu and Lee (2012) in terms of proper losses

The following are the results shown in Uematsu and Lee (2012).

Proposition 79 ((Uematsu and Lee, 2012, Theorem 3)) *Suppose $\ell(y, v) = \phi(yv)$ for some $\phi : \mathbb{R} \to \mathbb{R}_+$, where ϕ is differentiable, monotone decreasing, convex, and $\phi'(0) < 0$. For a given distribution $D = \langle M, \eta\rangle \in \Delta_{\mathcal{X}\times\{\pm 1\}}$, let*

$$s^* \in \mathcal{S}^*_{\mathrm{BR}}(D, \ell).$$

Then,

$$(\forall x, x' \in \mathcal{X})\, \eta(x) \neq \eta(x') \implies \mathrm{sign}(\mathrm{Diff}(s^*)(x, x')) = \mathrm{sign}(\eta(x) - \eta(x')).$$

If ϕ is strictly convex, then the above also holds when $\eta(x) = \eta(x')$.

Proposition 80 ((Uematsu and Lee, 2012, Theorem 7)) *Suppose $\ell(y, v) = \phi(yv)$ for some $\phi : \mathbb{R} \to \mathbb{R}_+$, where ϕ is differentiable, strictly monotone decreasing, convex, and $f : s \mapsto \frac{\phi'(-s)}{\phi'(s)}$ is strictly increasing. Given any $D = \langle M, \eta\rangle \in \Delta_{\mathcal{X}\times\{\pm 1\}}$,*

$$\mathcal{S}^*_{\mathrm{BR}}(D, \ell) \subseteq \mathcal{S}_{\mathrm{Decomp}}$$

if and only if $\phi'(-s)/\phi'(s) = e^{as}$ for some $a > 0$.

We show how to interpret these results in terms of proper composite losses. First, we show that the conditions of their theorems imply that ℓ is a proper composite margin loss.

Proposition 81 *Let ϕ be differentiable with $\phi'(0) < 0$, monotone decreasing, and strictly convex. Then, $\ell(y, v) = \phi(yv)$ is strictly proper composite.*

Proof Let ϕ meet the stated conditions. Since ϕ is convex and monotone decreasing with $\phi'(0) < 0$, then it must be true that

$$(\forall v \in \mathbb{R})(\phi'(v) \neq 0 \vee \phi'(-v) \neq 0).$$

Further, the function

$$f(v) \doteq \frac{\phi'(v)}{\phi'(-v)}$$

is continuous by differentiability of ϕ, and monotone by monotonicity and convexity of ϕ, since

$$f'(v) = \frac{1}{(\phi'(v))^2} \cdot \left(\phi'(-v)\phi''(v) + \phi''(-v)\phi'(v)\right) \leq 0.$$

When ϕ is *strictly* convex, f is strictly monotone because the numerator above cannot be 0. Thus, the conditions of Corollary 16 in Vernet et al. (2011) hold, and so ℓ is strictly proper composite. ■

Lemma 82 *Let ϕ be differentiable, strictly monotone decreasing, convex, and such that $f : s \mapsto \frac{\phi'(-s)}{\phi'(s)}$ is strictly increasing. Then, $\ell(y, v) = \phi(yv)$ is strictly proper composite.*

Proof The proof follows by the conditions of Corollary 16 in Vernet et al. (2011), as before, with invertibility $f : s \mapsto \frac{\phi'(-s)}{\phi'(s)}$ directly assumed rather than derived as a consequence of strict convexity. ■

By Lemma 73, the statement of Uematsu and Lee (2012, Theorem 3) is equivalent to saying that $\eta = g \circ s^*$ for some non-decreasing g when ϕ is convex, and g is strictly increasing when ϕ is strictly convex. Thus, this strictly convex part of the result is as per Corollary 48, except that the latter explicitly provides the form of the link function relating η and s^*.

The following shows that the conditions in their Theorem 7 imply that the inverse link function is of the form $\Psi^{-1}(v) = \frac{1}{1+e^{-av}}$, which means the result is a special case of Proposition 44 where ℓ is a margin loss.

Lemma 83 *Let $\ell \in \mathcal{L}_{\text{SPC}}(\Psi)$ be such that $\ell(y, v) = \phi(yv)$ for some differentiable $\phi : \mathbb{R} \to \mathbb{R}_+$. Then,*

$$(\forall a \in \mathbb{R} \setminus \{0\})(\forall v \in \mathbb{R})\, \Psi^{-1}(v) = \frac{1}{1+e^{-av}} \iff \phi'(-v)/\phi'(v) = e^{av}.$$

Proof The link function for a differentiable proper composite loss satisfies

$$\begin{aligned}(\forall v \in \mathbb{R})\, \Psi^{-1}(v) &= \frac{1}{1 - \frac{\ell'_1(v)}{\ell'_{-1}(v)}} \\ &= \frac{1}{1 + \frac{\phi'(v)}{\phi'(-v)}} \\ &= \frac{1}{1+e^{-av}}\end{aligned}$$

where the last line is true iff the asserted statement holds. ■

Appendix G. Empirical illustration of Corollary 45

We present an empirical illustration of the fact that Corollary 45 holds for an *asymmetric* proper composite loss. We work with a discrete distribution over N instances, where the instance i has probability M_i of being drawn, and has an associated probability η_i of having a positive label. A scorer s is then some vector in $\mathbb{R}^n$. Given a loss ℓ, the bipartite risk of the scorer s is

$$\begin{aligned}\mathbb{L}_{\text{BR}}(s; \boldsymbol{D}, \ell) &= \mathop{\mathbb{E}}_{\mathsf{X} \sim P, \mathsf{X}' \sim Q} \left[\ell_{\text{symm}}(s(\mathsf{X}) - s(\mathsf{X}'))\right] \\ &= \sum_{i=1}^{N} \sum_{j=1}^{N} \left[\eta_i \cdot (1-\eta_j) \cdot (\ell_1(s_i - s_j) + \ell_{-1}(s_j - s_i)\right] \\ &= \sum_{i=1}^{N} \sum_{j=1}^{N} \left[\eta_i \cdot (1-\eta_j) \cdot (\ell_1(\langle s, e_i - e_j \rangle) + \ell_{-1}(\langle s, e_j - e_i \rangle)\right],\end{aligned}$$

where e_i is the ith standard basis vector in $\mathbb{R}^n$. The Bayes-optimal risk is simply the minimiser of the above objective, and may be computed by numerical optimisation.

We performed 20 repetitions of the following experiment: for $N = 10$ instances, we draw $\eta_i \sim \text{Beta}(4, 3)$, $Z_i \sim \text{Beta}(6, 2)$, and set $M_i = Z_i / \sum_j Z_j$. We then scaled the η's to lie in $[0.01, 0.99]$, ensuring that the

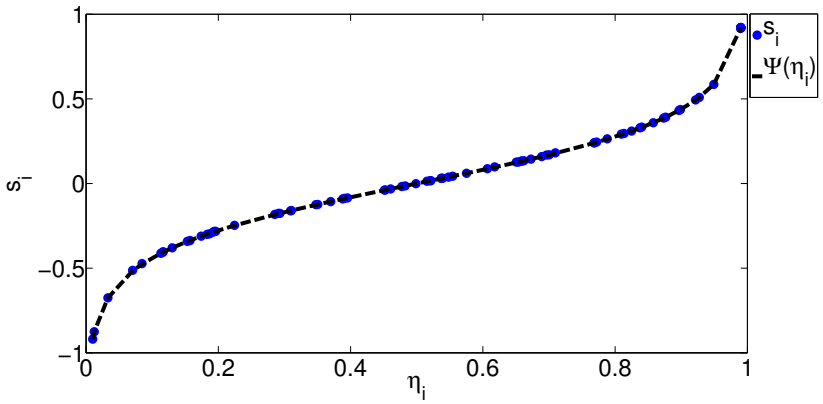


Figure 3: Results of 20 simulation trials to illustrate Bayes-optimal scorer (Corollary 45) for the case of an asymmetric loss. Here, the distribution D is varied across each trial, and the relationship between the (η, s^*) pairs across all trials is plotted. The relationship exactly matches that of $s^* = \Psi(\eta)$.

minimum and maximum values are attained. Given this distribution, we minimised the bipartite risk using L-BFGS, obtaining the Bayes-optimal scorer s^*. As the risk is invariant to translations, we transformed the solution so that its minimum value equals $\Psi(0.01)$ (thus agreeing with that of the expected optimal solution). We collected the corresponding pairs of (η_i, s_i^*) values for all 20 repetitions. We then plotted the graph of the resulting η values versus the s^* values. If s^* is a strictly monotone transform of η, then the plot will reflect this (as the different η values from the trials represent different sampling points of the domain of this function).

Figure 3 shows the results where ℓ is the asymmetric p-classification loss for $p = 2$,

$$\ell(v) = \left(\frac{1}{2} \cdot e^{2v}, e^{-v}\right)$$

We see that the relationship between the two is strictly monotone. Also shown on the graph is the plot of η versus $\Psi \circ \eta$, where $\Psi = \frac{1}{2}\sigma^{-1}$; this perfectly agrees with the observed s^* values, as predicted by the theory.

Appendix H. Empirical illustration of Corollary 48

We now present an empirical illustration of the facts that for a proper composite loss whose Bayes-optimal pair-scorer is non-decomposable, (a) the optimal univariate scorer is a strictly monotone transform of η, and (b) the transformation is distribution dependent. We repeated the setup of Appendix G, except that we worked with ℓ being the squared loss, $\ell(y, v) = (1 - yv)^2$, and the canonical boosting loss (Buja et al., 2005),

$$\ell(y, v) = \frac{yv}{2} + \sqrt{1 + \frac{v^2}{4}}.$$

Squared loss employs the identity link, while the canonical boosting loss uses the link $\Psi(\eta) = \frac{2\eta - 1}{\sqrt{\eta(1-\eta)}}$, and thus neither induce a decomposable pair-scorer according to Proposition 44.

Figure 4 shows that the relationship between η and s^* for these losses across multiple trials is *not* monotone, and significantly deviates from the optimal solution in the class-probability estimation setting, viz. $s^* = \Psi(\eta)$ for Ψ the identity mapping. This indicates that in general, the relationship between η and s^* is distribution dependent.

Figures 5 and 6 further studies the relationship between the two quantities for each individual trial. We see that, for a given trial (or equivalently for a given distribution), the relationship between η and s^* is strictly monotone, as expected. However, across different trials, it is evident that the precise monotone transformation is different.

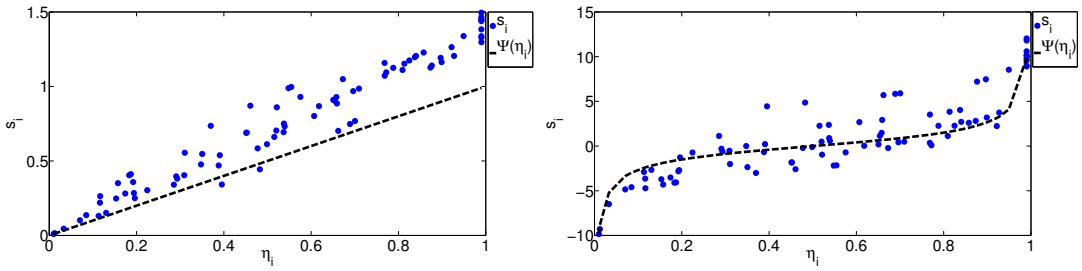


Figure 4: Results of 20 simulation trials to illustrate distribution dependent Bayes-optimal scorer (Proposition 47) for the case of squared and canonical boosting losses. Here, the distribution D is varied across each trial, and the relationship between the (η, s^*) pairs across all trials is plotted.

Appendix I. Empirical illustration of optimal p-norm push pair-scorer

We now present an empirical illustration of the fact that for a general proper composite loss, (a) the optimal p-norm push pair-scorer is a strictly monotone transform of η_{Pair}, and (b) the transformation is distribution dependent. We repeated the setup of Appendix G, except that we worked with the p-norm push risk for $p = 4$, with ℓ being logistic loss, and considered $n = 5$ to reduce the number of η_{Pair} values.

Figure 7 shows the relationship between η_{Pair} and s^*_{Pair} across multiple trials is *not* monotone, and significantly deviates from the optimal solution in the class-probability estimation setting, viz. $s^* = \Psi(\eta)$ for $\Psi = \frac{1}{p}\sigma^{-1}$. This indicates that in general, the relationship between η_{Pair} and s^*_{Pair} is distribution dependent. Figure 8 further studies the relationship between the two quantities for each individual trial. We see that, for a given trial (or equivalently for a given distribution), the relationship between η_{Pair} and s^*_{Pair} is strictly monotone, as expected.

Appendix J. Empirical illustration of optimal p-norm push univariate scorer

We now present an empirical illustration of the fact that for a general proper composite loss, the optimal p-norm push univariate scorer is a distribution dependent transform of η. We repeat the setup of Appendix G, except that we worked with the p-norm push risk for $p = 4$, with ℓ being logistic loss, and considered $n = 5$ to reduce the number of η_{Pair} values.

Figure 9 shows the relationship between η and s^* across multiple trials is *not* monotone, and significantly deviates from the optimal solution in the class-probability estimation setting, viz. $s^* = \Psi(\eta)$ for $\Psi = \frac{1}{p}\sigma^{-1}$.

Appendix K. Illustration of Hand's representation and proper loss equivalence

We illustrate empirically that the AUC for certain calibrated scorers is equivalent to a suitable risk with respect to a proper loss, owing to Hand's representation (Equation 34). For a fixed n, we consider a finite instance space $\mathcal{X} = \{0, 1/n, 2/n, \ldots, 1\}$. We consider a distribution D over $\mathcal{X}$ where the marginal M is uniform and the class-probability function η is of a pre-specified form. Specifically, we considered $\mathbb{P}[\eta(\mathsf{X}) = c] \propto w(c)$, where $w(c)$ is the weight function corresponding to a proper loss. We considered two choices of w: $w(c) = 1$, corresponding to square loss, and $w(c) = 1/\sqrt{c \cdot (1-c)}$, corresponding to the arcsin loss of Buja et al. (2005, Section 11).

For a given n, we then computed the AUC of the scorer $s^* = \eta$, and compared it to the corresponding proper loss of the scorer. Hand's representation (Equation 34) suggests that as $n \to \infty$, so that the distribution of the scores is exactly the weight of the proper loss, the two will be equivalent.

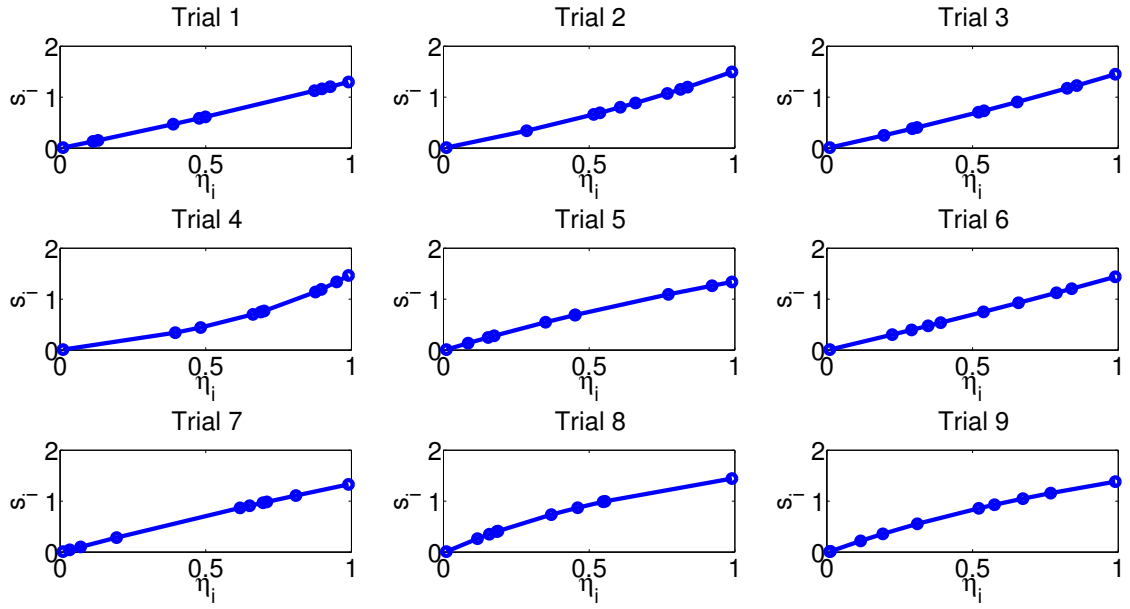


Figure 5: Results of 9 simulation trials to illustrate order preserving Bayes-optimal scorer (Corollary 48) for the case of squared loss. Here, the distribution D is varied across each trial, and each panel represents the relationship between η and s^* for a *specific* trial.

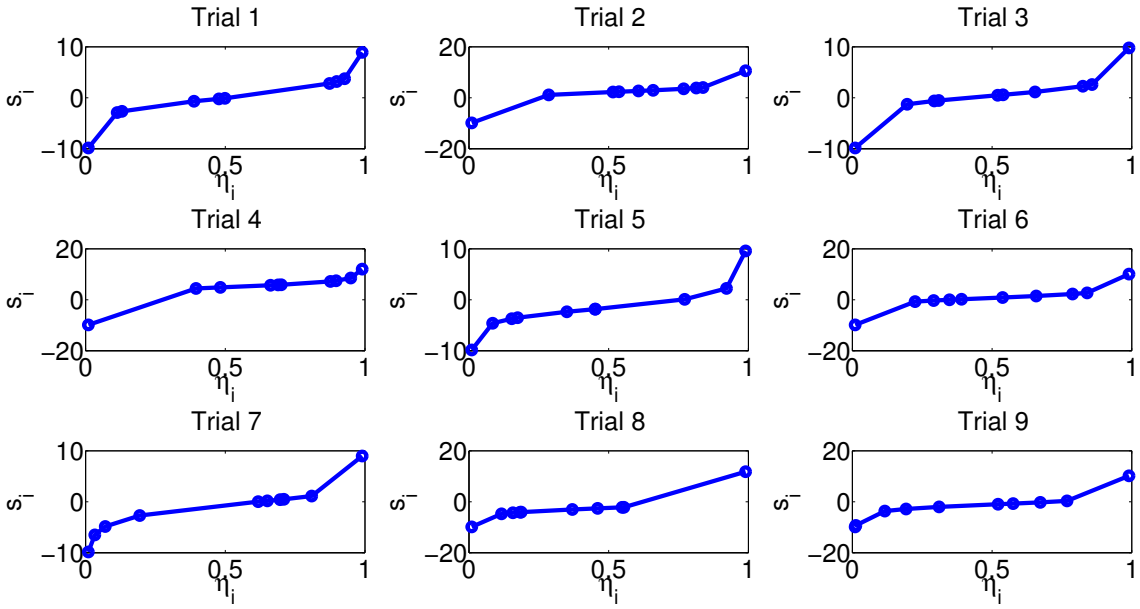


Figure 6: Results of 9 simulation trials to illustrate order preserving Bayes-optimal scorer (Corollary 48) for the case of canonical boosting loss. Here, the distribution D is varied across each trial, and each panel represents the relationship between η and s^* for a *specific* trial.

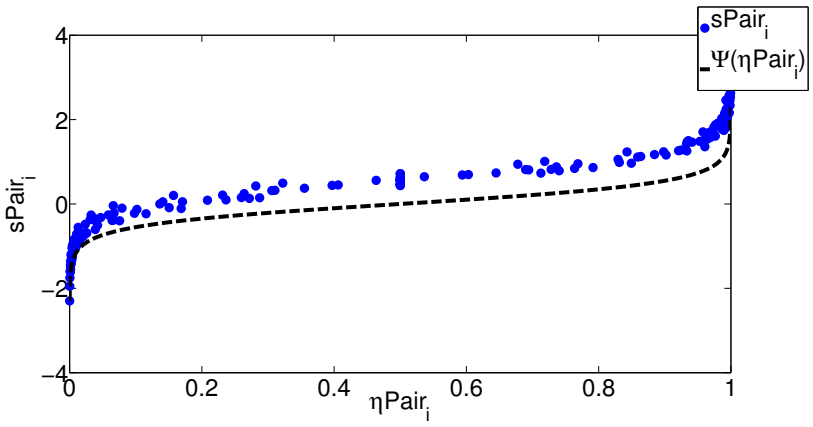


Figure 7: Results of 9 simulation trials to illustrate the relationship between η_{Pair} and s^*_{Pair} for p-norm push with logistic loss. Here, the distribution D is varied across each trial, and the relationship between the $(\eta_{\text{Pair}}, s^*_{\text{Pair}})$ pairs across all trials is plotted.

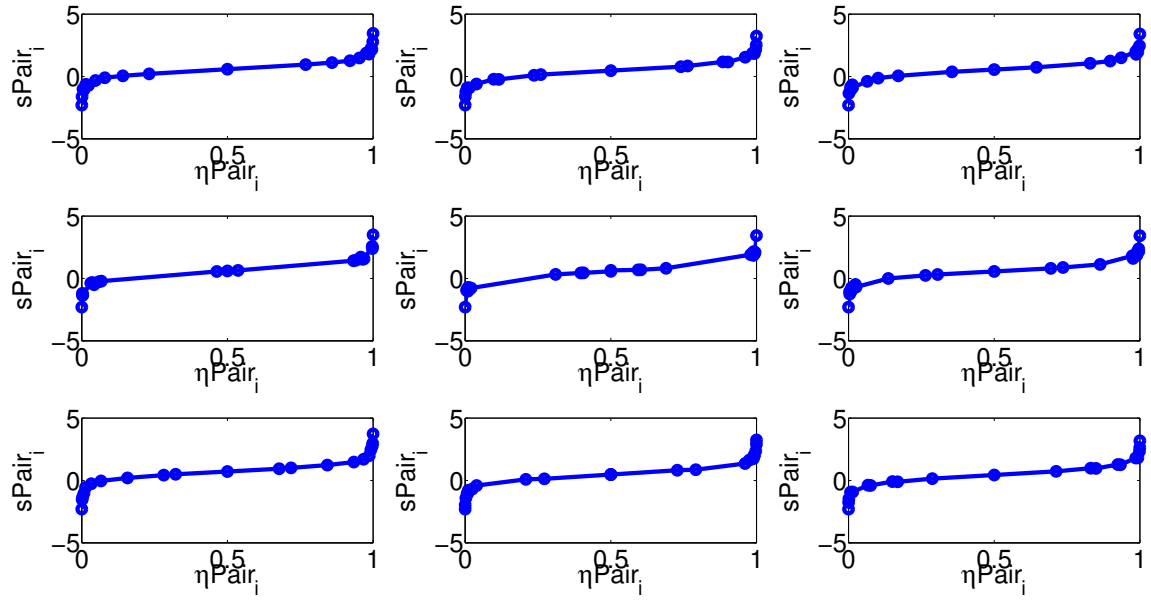


Figure 8: Results of 9 simulation trials to illustrate the distribution dependent relationship between η_{Pair} and s^*_{Pair} for p-norm push with logistic loss. Here, the distribution D is varied across each trial, and each panel represents the relationship between η_{Pair} and s^*_{Pair} for a *specific* trial.

Figure 10 shows that for a large n, one minus the AUC and the appropriate proper risk converge. The results compare the two over 100 trials, where each trial corresponds to a different random draw of η from the distribution with density given by the appropriately normalised weight $w(\cdot)$.

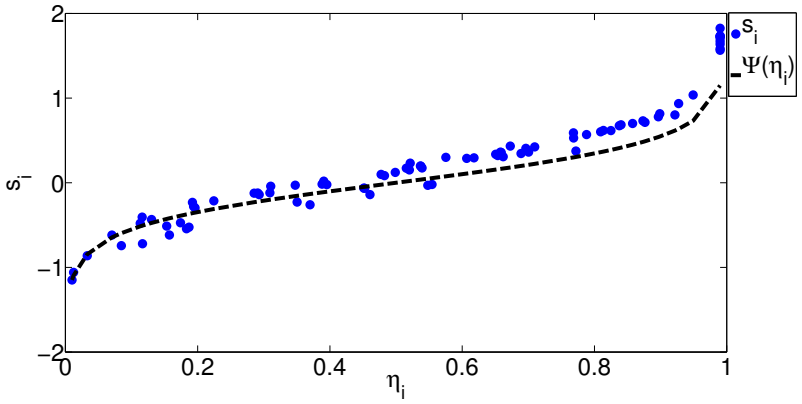


Figure 9: Results of 20 simulation trials to illustrate the relationship between η and s^* for p-norm push with logistic loss. Here, the η_i and M_i values were varied across each trial, and each panel represents the relationship between η and s^* for a *specific* trial.

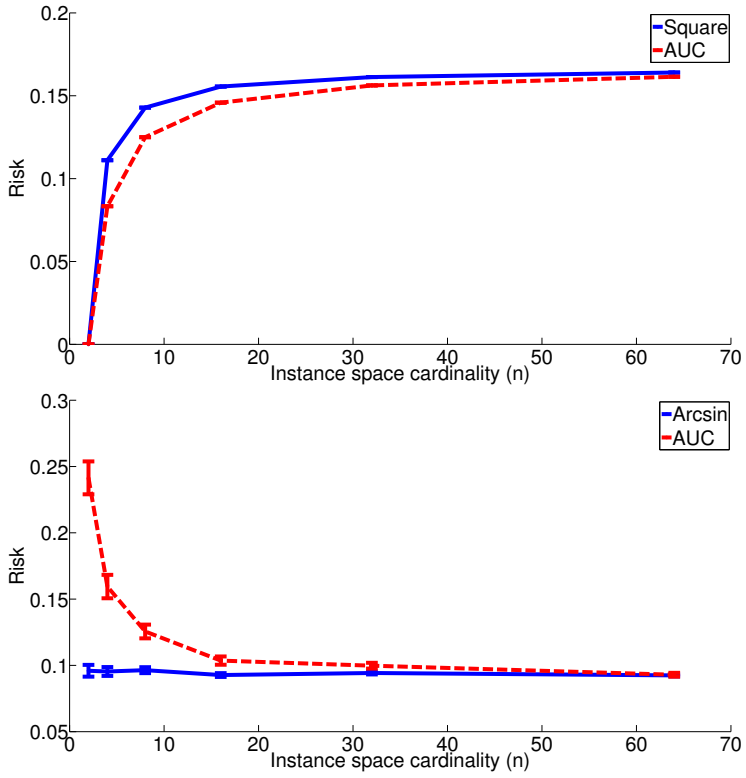


Figure 10: Results of 100 simulation trials to illustrate the relationship between the AUC and a proper risk, for different choices of distribution on the underlying optimal scorer $s^* = \eta$ on an instance space with n elements.

References

Shivani Agarwal. The infinite push: A new support vector ranking algorithm that directly optimizes accuracy at the absolute top of the list. In *SIAM International Conference on Data Mining (SDM)*, pages 839–850, 2011.

Shivani Agarwal. Surrogate regret bounds for bipartite ranking via strongly proper losses. *Journal of Machine Learning Research*, 15:1653–1674, 2014.

Shivani Agarwal and Partha Niyogi. Stability and generalization of bipartite ranking algorithms. In *Conference on Learning Theory (COLT)*, pages 32–47, Berlin, Heidelberg, 2005.

Shivani Agarwal, Thore Graepel, Ralf Herbrich, Sariel Har-Peled, and Dan Roth. Generalization bounds for the area under the ROC curve. *Journal of Machine Learning Research*, 6:393–425, December 2005.

Alan Agresti. *Analysis of ordinal categorical data*. Wiley Series in Probability and Statistics, 1984.

Nir Ailon and Mehryar Mohri. An efficient reduction of ranking to classification. *CoRR*, abs/0710.2889, 2007.

Miriam Ayer, Hugh D. Brunk, George M. Ewing, William T. Reid, and Edward Silverman. An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, 26 (4):641–647, 12 1955.

Bernard De Baets, Hans De Meyer, and Bart De Schuymer. Transitive comparison of random variables. In Erich Petr Klement and Radko Mesiar, editors, *Logical, Algebraic, Analytic and Probabilistic Aspects of Triangular Norms*, pages 415–442. Elsevier, Amsterdam, 2005.

Maria-Florina Balcan, Nikhil Bansal, Alina Beygelzimer, Don Coppersmith, John Langford, and Gregory B . Sorkin. Robust reductions from ranking to classification. *Machine Learning*, 72(1-2):139–153, 2008.

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Kendrick Boyd, Kevin H. Eng, and C. David Page. Area under the precision-recall curve: Point estimates and confidence intervals. In Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Železný, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 8190 of *Lecture Notes in Computer Science*, pages 451–466. Springer Berlin Heidelberg, 2013.

Stephen P. Boyd, Corinna Cortes, Mehryar Mohri, and Ana Radovanovic. Accuracy at the top. In *Advances In Neural Information Processing Systems (NIPS)*, pages 962–970, 2012.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs I: The method of paired comparisons. *Biometrika*, 39(3/4):pp.324–345, 1952.

Lev M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200 – 217, 1967.

Douglas S. Bridges and Ghanshyam B. Mehta. *Representations of preference orderings*. Lecture notes in economics and mathematical systems. Springer, 1995.

Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. www-stat.wharton.upenn.edu/~buja/PAPERS/paper-proper-scoring.pdf, 2005. Unpublished manuscript.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *International Conference on Machine learning (ICML)*, pages 89–96, 2005.

Soumen Chakrabarti, Rajiv Khanna, Uma Sawant, and Chiru Bhattacharyya. Structured learning for non-smooth ranking losses. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 88–96, 2008.

Philip K. Chan and Salvatore J. Stolfo. Learning with non-uniform class and cost distributions: Effects and a multi-classifier approach. In *KDD 1998 Workshop on Distributed Data Mining*, pages 1–9, 1998.

Stéphan Clémençon and Nicolas Vayatis. Ranking the best instances. *Journal of Machine Learning Research*, 8:2671–2699, December 2007.

Stéphan Clémençon and Nicolas Vayatis. Empirical performance maximization for linear rank statistics. In *Advances in Neural Information Processing Systems (NIPS)*, pages 305–312, 2008.

Stéphan Clémençon and Nicolas Vayatis. Nonparametric estimation of the precision-recall curve. In *International Conference on Machine Learning (ICML)*, pages 185–192, 2009a.

Stéphan Clémençon and Nicolas Vayatis. Adaptive estimation of the optimal ROC curve and a bipartite ranking algorithm. In *International Conference on Algorithmic Learning Theory (ALT)*, pages 216–231, 2009b.

Stéphan Clémençon, Gábor Lugosi, and Nicolas Vayatis. Ranking and Empirical Minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874, April 2008.

Stéphan Clémençon, Marine Depecker, and Nicolas Vayatis. AUC optimization and the two-sample problem. In *Advances in Neural Information Processing Systems (NIPS)*, pages 360–368, 2009.

Stéphan Clémençon and Nicolas Vayatis. Tree-based ranking methods. *IEEE Transactions on Information Theory*, 55(9):4316–4336, Sept 2009.

Stéphan Clémençon, Sylvain Robbiano, and Nicolas Vayatis. Ranking data with ordinal labels: optimality and pairwise aggregation. *Machine Learning*, 91(1):67–104, 2013.

William W. Cohen, Robert E. Schapire, and Yoram Singer. Learning to order things. *Journal of Artificial Intelligence Research*, 10(1):243–270, May 1999.

Corinna Cortes and Mehryar Mohri. AUC optimization vs. error rate minimization. In *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, 2003.

David Cossock and Tong Zhang. Statistical analysis of Bayes optimal subset ranking. *IEEE Transactions on Information Theory*, 54(11):5140–5154, November 2008.

Koby Crammer and Yoram Singer. Pranking with ranking. In *Advances in Neural Information Processing Systems (NIPS)*, pages 641–647. MIT Press, 2001.

Imre Csiszár. Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *A Magyar Tudományos Akadémia Matematikai és Fizikai Tudományok Osztályának Közleményei*, 8:85–108, 1963.

Mark A. Davenport, Richard G. Baraniuk, and Clayton D. Scott. Tuning support vector machines for minimax and Neyman-Pearson classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32 (10):1888–1898, October 2010.

Gerard Debreu. Representation of a preference ordering by a numerical function. In R. M. Thrall, C. H. Coombs, and R. L. Davis, editors, *Decision Processes*, pages 159–65. Wiley, New York, 1954.

Gerard Debreu. Continuity properties of Paretian utility. *International Economic Review*, 5(3):pp.285–293, 1964.

Morris H. DeGroot and Stephen E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32(1/2):pp.12–22, 1983.

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

Lori E. Dodd and Margaret S. Pepe. Partial AUC estimation and regression. *Biometrics*, 59(3):pp.614–623, 2003.

James P. Egan. *Signal Detection Theory and ROC Analysis*. Series in Cognition and Perception. Academic Press, 1975.

Samuel Eilenberg. Ordered topological spaces. *American Journal of Mathematics*, 63(1):pp.39–45, 1941.

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 213–220, 2008.

Şeyda Ertekin and Cynthia Rudin. On equivalence relationships between classification and ranking algorithms. *Journal of Machine Learning Research*, 12:2905–2929, Oct 2011.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, June 2006.

Tom Fawcett and Alexandru Niculescu-Mizil. PAV and the ROC convex hull. *Machine Learning*, 68(1): 97–106, 2007.

Cèsar Ferri, Peter Flach, and Athmane Senad. Modifying ROC curves to incorporate predicted probabilities. In *International Conference on Machine Learning (ICML) Workshop on ROC Analysis in ML*, 2005.

Peter Flach, Josè Hernândez-Orallo, and Cësar Ferri. A coherent interpretation of AUC as a measure of aggregated classification performance. In *International Conference on Machine Learning (ICML)*, June 2011.

Peter A. Flach. ROC analysis. In Claude Sammut and Geoffrey I. Webb, editors, *Encyclopedia of Machine Learning*, pages 869–875. Springer, 2010.

Gerald B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Wiley Interscience, New York, 1999.

Yoav Freund, Raj Iyer, Robert E. Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4:933–969, December 2003.

Johannes Fürnkranz and Eyke Hüllermeier. *Preference Learning*. Springer-Verlag, 1st edition, 2010.

Wei Gao and Zhi-Hua Zhou. On the consistency of AUC optimization. *CoRR*, abs/1208.0645, 2012.

Wei Gao and Zhi-Hua Zhou. On the consistency of AUC pairwise optimization. In *International Joint Conference on Artificial Intelligence*, 2015.

Gilles Gasso, Aristidis Pappaioannou, Marina Spivak, and Léon Bottou. Batch and online learning algorithms for nonconvex Neyman-Pearson classification. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):28:1–28:19, May 2011.

Izrail M. Gelfand and Sergei V. Fomin. *Calculus of Variations*. Dover, 2000.

Mariano Giaquinta and Stefan Hildebrandt. *Calculus of Variations I: The Lagrangian formalism*. Springer-Verlag, Berlin, 2nd edition, 2004.

Tilmann Gneiting and Matthias Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1(1):125–151, 2014.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, March 2007.

David J. Hand. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, 77(1):103–123, October 2009.

David J. Hand and Robert J. Till. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, 45(2):171–186, 2001.

James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36, 1982.

Ralf Herbrich, Thore Graepel, Peter Bollmann-Sdorra, and Klaus Obermayer. Learning Preference Relations for Information Retrieval. In *AAAI Workshop Text Categorization and Machine Learning*, pages 80–84, Madison, 1998.

Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Large margin rank boundaries for ordinal regression. In A.J. Smola, P.L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 115–132. MIT Press, Cambridge, MA, 2000.

José Hernández-Orallo, Peter Flach, and Cèsar Ferri. A Unified View of Performance Metrics: Translating Threshold Choice into Expected Classification Loss. *Journal of Machine Learning Research*, 13:2813–2868, October 2012.

Ralph D. Hippenstiel. *Detection Theory: Applications and Digital Signal Processing*. CRC Press, University of Texas at Tyler, USA, 2001.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, October 2002.

Thorsten Joachims. Optimizing search engines using clickthrough data. In *ACM SIGKDD International Conference on Knowledge discovery and data mining (KDD)*, pages 133–142, 2002.

Palaniappan Kannappan. *Functional equations and inequalities with applications.* Springer, New York, 2009.

Donald E. Knuth. Two notes on notation. *American Mathematical Monthly*, 99(5):403–422, May 1992.

Wojciech Kotlowski, Krzysztof Dembczynski, and Eyke Hüllermeier. Bipartite ranking through minimization of univariate loss. In *International Conference on Machine Learning (ICML)*, pages 1113–1120, 2011.

David H. Krantz. Rational distance functions for multidimensional scaling. *Journal of Mathematical Psychology*, 4:226 – 245, 1967.

Wojtek J. Krzanowski and David J. Hand. *ROC Curves for Continuous Data*. Chapman & Hall/CRC, 1st edition, 2009.

John Langford and Bianca Zadrozny. Estimating class membership probabilities using classifier learners. In *International Workshop on Artificial Intelligence and Statistics (AISTATS)*, 2005.

Erik L. Lehmann and Joseph P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005.

Ling Li and Hsuan-Tien Lin. Ordinal regression by extended binary classification. In *Advances In Neural Information Processing Systems (NIPS)*, pages 865–872, 2006.

Nan Li, Rong Jin, and Zhi-Hua Zhou. Top rank optimization in linear time. *Advances in Neural Information Processing Systems*, pages 1–9, 2014.

Charles X. Ling and Chenghui Li. Data mining for direct marketing: Problems and solutions. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 217–225, 1998.

Tie-Yan Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3 (3):225–331, March 2009.

Robert Duncan Luce. *Individual Choice Behavior*. Wiley, New York, 1959.

Robert Duncan Luce and Patrick Suppes. Preference, Utility, and Subjective Probability. *Handbook of mathematical psychology*, 3(171):249–410, 1965.

Henry B. Mann and Donald R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 03 1947.

Hamed Masnadi-Shirazi and Nuno Vasconcelos. Variable margin losses for classifier design. In *Advances In Neural Information Processing Systems (NIPS)*, pages 1576–1584, 2010.

Donna Katzman McClish. Analyzing a portion of the ROC curve. *Medical Decision Making*, 9(3):190–195, 1989.

Aditya Krishna Menon and Robert C. Williamson. Bayes-optimal scorers for bipartite ranking. In *Conference on Learning Theory (COLT)*, 2014.

Aditya Krishna Menon, Brendan van Rooyen, Cheng Soon Ong, and Robert C. Williamson. Learning from corrupted binary labels via class-probability estimation. In *International Conference on Machine Learning (ICML)*, pages 125–134, 2015.

Harikrishna Narasimhan and Shivani Agarwal. SVM_{pAUC}tight: a new support vector method for optimizing partial AUC based on a tight convex upper bound. In *ACM SIGKDD International Conference on Knowledge discovery and data mining (KDD)*, pages 167–175, 2013a.

Harikrishna Narasimhan and Shivani Agarwal. On the relationship between binary classification, bipartite ranking, and binary class probability estimation. In *Advances In Neural Information Processing Systems (NIPS)*, pages 2913–2921, 2013b.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, New York, 2nd edition, 2006.

Ferdinand Österreicher and Igor Vajda. Statistical information and discrimination. *IEEE Transactions on Information Theory*, 39(3):1036–1039, 1993.

John R. Platt. Strong inference. *Science*, 146(3642):347–353, October 1962.

Foster Provost and Tom Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42 (3):203–231, March 2001.

Mark D. Reid and Robert C. Williamson. Surrogate regret bounds for proper losses. In *International Conference on Machine Learning (ICML)*, pages 897–904, 2009.

Mark D. Reid and Robert C. Williamson. Composite binary losses. *Journal of Machine Learning Research*, 11:2387–2422, December 2010.

Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12:731–817, Mar 2011.

Fred S. Roberts. *Measurement theory with Applications to Decision Making, Utility, and the Social Sciences*, volume 7 of *Encyclopedia of Mathematics and Its Applications*. Addison-Wesley, Reading, MA, 1984.

Cynthia Rudin. The p-norm push: A simple convex ranking algorithm that concentrates at the top of the list. *Journal of Machine Learning Research*, 10:2233–2271, December 2009.

Walter Rudin. *Functional Analysis*. McGraw-Hill Book Co., New York, 2nd edition, 1973. McGraw-Hill Series in Higher Mathematics.

Mark J. Schervish. A general method for comparing probability assessors. *The Annals of Statistics*, 17(4): 1856–1879, 12 1989.

Clayton Scott and Mark Davenport. Regression level set estimation via cost sensitive classification. *IEEE Transactions on Signal Processing*, 55:2752–2757, 2007.

Martin J. J. Scott, Mahesan Niranjan, and Richard W. Prager. Realisable classifiers: Improving operating performance on variable cost problems. In *British Machine Vision Conference*, pages 304–315, 1998.

Sundararajan Sellamanickam, Priyanka Garg, and Sathiya Keerthi Selvaraj. A pairwise ranking based approach to learning with positive and unlabeled examples. In *ACM International Conference on Information and Knowledge Management (CIKM)*, pages 663–672, 2011.

Amnon Shashua and Anat Levin. Ranking with large margin principle: Two approaches. In *Advances In Neural Information Processing Systems (NIPS)*, pages 937–944, 2002.

Emir H. Shuford Jr., Arthur Albert, and H. Edward Massengill. Admissible probability measurement procedures. *Psychometrika*, 31(2):125–145, 1966.

Ingo Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26(2): 225–287, 2007.

Roy L. Streit. A neural network for optimum Neyman-Pearson classification. In *International Joint Conference on Neural Networks (IJCNN)*, volume 1, pages 685–690, 1990.

Robert S. Strichartz. *A Guide to Distribution Theory and Fourier Transforms*. Studies in Advanced Mathematics. CRC Press, Boca Raton, FL, 1994.

S. Joshua Swamidass, Chloé-Agathe Azencott, Kenny Daily, and Pierre Baldi. A CROC stronger than ROC. *Bioinformatics*, 26(10):1348–1356, May 2010.

Zbigniew Świtalski. General transitivity conditions for fuzzy reciprocal preference matrices. *Fuzzy Sets and Systems*, 137(1):85–100, 2003.

Erik N. Torgersen. *Comparison of Statistical Experiments*. Cambridge University Press, 1991.

John L. Troutman. *Variational Calculus and Optimal Control: Optimization with Elementary Convexity*. Undergraduate Texts in Mathematics. Springer, 1996.

Kazuki Uematsu and Yoonkyung Lee. On theoretically optimal ranking functions in bipartite ranking. http://www.stat.osu.edu/~yklee/mss/bipartrank.rev.pdf, 2012. Unpublished manuscript.

Elodie Vernet, Mark D. Reid, and Robert C. Williamson. Composite multiclass losses. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1224–1232, 2011.

Ellen M. Voorhees. The TREC question answering track. *Natural Language Engineering*, 7(4):361–378, December 2001. ISSN 1351-3249.

Shaomin Wu and Peter Flach. A scored AUC metric for classifier evaluation and selection. In *International Conference on Machine Learning (ICML) Workshop on ROC Analysis in ML*, 2005.

Jingdong Xie and Carey E. Priebe. A weighted generalization of the Mann-Whitney-Wilcoxon statistic. *Journal of Statistical Planning and Inference*, 102(2):441 – 466, 2002.

Yisong Yue, Thomas Finley, Filip Radlinski, and Thorsten Joachims. A support vector method for optimizing average precision. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 271–278, 2007.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32:56–134, March 2004.